DOCUMENT RESUME

ED 025 288 52 LI 001 241

By- Schultz, Claire K.; And Others

Evaluation of Indexing by Group Consensus. Final Report.

Institute for the Advancement of Medical Communication, Philadelphia, Pa.
Spons Agency- Office of Education (DHEW), Washington, D.C. Bureau of Research.
Bureau No- BR-7-0622
Pub Date 30 Aug 68
Contract- OEC-1-7-070622-3890
Note- 46p.
EDRS Price MF-$0.25 HC-$2.40
Descriptors- Analysis of Variance, Comparative Analysis, *Evaluation Criteria, *Evaluation Techniques, *Groups, *Indexing, *Information Retrieval, Performance Factors, Reliability, Standards
Identifiers- *Criterion Group Method

This study was designed to explore the practicality, flexibility, reliability, and sensitivity of the Criterion Group Method of measuring the effectiveness and efficiency of indexing, a method using a criterion group to set the standard for "ideal" indexing. These major variables were examined: (1) size of document sample, (2) size of Criterion Group, (3) instructions to indexers and use of a vocabulary guide, (4) three methods of editing raw indicia to make terms comparable, and (5) two methods of weighting indexers' scores. Scores earned by a set of eight professional indexers, by individual authors of the test documents and, in some cases, scores for title sets or medical students' indexing were compared within selected treatments to measure the extent to which the detectability of differences was achieved by each treatment. A two-way analysis of variance was used to relate reliability of test scores to document sample size and criterion group size. From the results of these studies of the methodologic variables, it was concluded that the criterion group method of evaluating indexing can be a practical yardstick for a wide variety of managerial, research, and educational uses. Appendixes include the rationale of the method, a literature review, information on materials employed and subjects participating in study trials, and details on manual and computer implementation. (Author/JB)

BR 7-0622
PA-52
LI 001241

FINAL REPORT

PROJECT NO. 7-0622

CONTRACT NO. OEC 1-7-070622-3890

EVALUATION OF INDEXING BY GROUP CONSENSUS

AUGUST 30, 1968

ED025288

U. S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

OFFICE OF EDUCATION
BUREAU OF RESEARCH

LI 001241

FINAL REPORT

PROJECT NO. 7-0622

CONTRACT NO. OEC 1-7-070622-3890

EVALUATION OF INDEXING BY GROUP CONSENSUS

CLAIRE K. SCHULTZ, PETER B. HENDERSON, AND RICHARD H. ORR

INSTITUTE FOR ADVANCEMENT OF MEDICAL COMMUNICATION

PHILADELPHIA, PA. 19104

AUGUST 30, 1968

THE RESEARCH REPORTED HEREIN WAS PERFORMED PURSUANT TO A CONTRACT OEC-1-7-070622-3890 WITH THE OFFICE OF EDUCATION, U. S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE. CONTRACTORS UNDERTAKING SUCH PROJECTS UNDER GOVERNMENT SPONSORSHIP ARE ENCOURAGED TO EXPRESS FREELY THEIR PROFESSIONAL JUDGMENT IN THE CONDUCT OF THE PROJECT. POINTS OF VIEW OR OPINIONS STATED DO NOT, THEREFORE, NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

OFFICE OF EDUCATION
BUREAU OF RESEARCH

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

ACKNOWLEDGEMENTS

MRS. RITA PORRECA, WORKING UNDER THE DIRECTION OF DR. MYRNA GOPNIK, WROTE THE COMPUTER PROGRAM FOR A CONTEXT-DEPENDENT METHOD OF THESAURUS PROCESSING. DR. ANDREW BAGGALEY, DEPT. OF EDUCATION, UNIVERSITY OF PENNSYLVANIA, SERVED AS STATISTICAL CONSULTANT TO THE PROJECT.

WE ARE MOST GRATEFUL TO THE SCIENTISTS AND PROFESSIONAL INDEXERS WHO GAVE THEIR TIME, AND TO THE STUDENTS WHO WORKED AT A NOMINAL FEE, TO GENERATE THE DATA THAT MADE THIS STUDY POSSIBLE.

TABLE OF CONTENTS

	Page
INTRODUCTION	1
Aim of study	1
Methodologic desiderata	1
Organization of report	2
ESSENTIALS OF METHOD	3
Selecting the document sample	3
Selecting and constructing the criterion group	3
Instructing test indexers	4
Establishing criterion and test sets	4
Weighting the criterion sets	5
Scoring the test sets	5
FINDINGS ON METHODOLOGIC VARIABLES	7
Variable 1 - size of document sample	7
Variable 2 - size of criterion group	9
Variable 3 - instructions to test indexers	11
Variable 4 - procedures for editing criterion and test sets	13
Variable 5 - weighting scheme for scoring	16
Variable 6 - confounding before scoring	17
CONCLUSIONS	17
APPENDIX A	
RATIONALE OF METHOD AND LITERATURE REVIEW	A 1-12
APPENDIX B	
MATERIALS EMPLOYED IN STUDY TRIALS	B 1-3
APPENDIX C	
SUBJECTS PARTICIPATING IN STUDY TRIALS	C 1-2
APPENDIX D	
PROCEDURES FOR MANUAL IMPLEMENTATION	D 1-2
APPENDIX E	
COMPUTER IMPLEMENTATION	E 1-4

SUMMARY

THE CRITERION GROUP METHOD TESTS THE EFFECTIVENESS AND EFFICIENCY OF TEST INDEXING SETS, USING A CRITERION GROUP TO SET THE STANDARD FOR "IDEAL" INDEXING. THE CRITERION GROUP FOR A PARTICULAR APPLICATION IS CHOSEN BY THE TEST ADMINISTRATOR, CONSISTENT WITH HIS OWN CONCEPT OF WHO REPRESENTS THIS "IDEAL". MATCHING TEST SETS OF INDEXING TERMS WITH THE CRITERION SET YIELDS AS MANY DEGREES OF MATCH AS THERE ARE MEMBERS OF THE CRITERION GROUP (REFERRED TO AS CONCENSUS NUMBER).

THIS STUDY WAS DESIGNED TO EXPLORE THE PRACTICALITY, FLEXIBILITY, RELIABILITY, AND SENSITIVITY OF THE METHOD. TO DO THIS, IT EXAMINES THE MAJOR VARIABLES: (1) SIZE OF THE DOCUMENT SAMPLE, (2) SIZE OF THE CRITERION GROUP, (3) EFFECT OF VARIOUS INSTRUCTIONS TO INDEXERS AND USE OF A VOCABULARY GUIDE, (4) EFFECTS OF THREE METHODS OF EDITING RAW INDICIA TO MAKE TERMS COMPARABLE, AND (5) TWO ALTERNATIVE METHODS OF WEIGHTING INDEXERS' SCORES.

SCORES EARNED BY A SET OF EIGHT PROFESSIONAL INDEXERS, BY INDIVIDUAL AUTHORS OF THE TEST DOCUMENTS, AND IN SOME CASES SCORES FOR TITLE SETS OR MEDICAL STUDENTS' INDEXING, WERE COMPARED WITHIN SELECTED TREATMENTS TO MEASURE THE EXTENT TO WHICH THE DETECTABILITY OF DIFFERENCES WAS ACHIEVED BY EACH TREATMENT. A TWO-WAY ANALYSIS OF VARIANCE WAS USED TO RELATE RELIABILITY OF TEST SCORES TO DOCUMENT SAMPLE SIZE AND CRITERION GROUP SIZE.

RESULTS WITH REGARD TO PRACTICALITY SHOW THAT "INDICATIVE" TESTS (ALLOWING CONFIDENCE LIMITS OF $\pm$ 10 POINTS) AT THE 80% LEVEL OF CONFIDENCE CAN BE MADE WITH DOCUMENT SAMPLES AS SMALL AS 10 AND CRITERION GROUPS AS SMALL AS 4; 95% CONFIDENCE REQUIRES COMPARABLE VALUES OF 20 DOCUMENTS AND 9 CRITERION GROUP MEMBERS. IT IS POSSIBLE TO CONDUCT TESTS WITH ONLY A FEW "MECHANICAL" INSTRUCTIONS TO INDEXERS, NO VOCABULARY GUIDE, NO EDITING AND NO WEIGHTING DURING SCORING OTHER THAN USE OF THE CONCENSUS NUMBER EVEN THOUGH FROM THE STANDPOINT OF SENSITIVITY, THE METHOD CAN DETECT DIFFERENCES IN SCORES DUE TO EDITING METHOD, INSTRUCTIONS TO INDEXERS, USE OF A VOCABULARY GUIDE, OR WEIGHTING METHOD, SHOULD SUCH DETECTION BE DESIRABLE. THE METHOD IS FLEXIBLE IN THAT IT HAS BEEN SHOWN TO LEAVE AS OPTIONS VARIABLES SUCH AS METHOD OF INSTRUCTING INDEXERS, METHOD OF EDITING AND METHOD OF WEIGHTING. RELIABILITY IS PRIMARILY DEPENDENT ON THE SIZE OF DOCUMENT SAMPLE AND CRITERION GROUP, AS IS DEMONSTRATED GRAPHICALLY IN THE PAPER. FACE VALIDITY, OR INTUITIVE FEEL FOR THE MEANING OF TEST RESULTS, IS ENHANCED BY THE FACT THAT DIFFERENCES IN SCORES CAN BE EQUATED WITH DIFFERENCES IN THE INTERNATIONALLY KNOWN MEASURE OF "RECALL", AND SCORE DIVIDED BY THE NUMBER OF TERMS IN THE INDEXING SET YIELDS A RESULT SOMEWHAT ANALOGOUS TO "PRECISION". PERCENT MAXIMAL SCORE CAN ALSO BE EASIER TO ENVISION AS REFLECTING EFFECTIVENESS OF TEXT INDEXING SETS ON A 0-100 SCALE, WITH DIFFERENCES OF FROM 6-8 POINTS REPRESENTING SIGNIFICANCE, WHEN SIGNED-RANK TESTS ARE APPLIED.

INTRODUCTION

Aim of Study

Several years ago, for a special research application, we developed a method for measuring the "quality" or "effectiveness", of indexing. * At that time we felt this method could be adapted for a wide range of managerial, educational, and research applications and, since it had several important advantages over other methods,# it might fill the critical need for a practical yardstick to evaluate indexing and subject cataloging. However, there were a number of questions to be answered before one could be certain that the method met the demanding requirements for such a yardstick. The present study was undertaken to explore these questions.

Methodologic Desiderata

For a truly general method, applicable to many types of indexing and subject cataloging and suitable for serving a wide range of purposes, certain methodologic characteristics would seem to be either essential or highly desirable. First, the method should have "face" validity in the eyes of those who will use the resulting measurements; and since individuals have varying concepts of what constitutes "ideal" indexing, the method should allow one the option of choosing a criterion concept that reflects his own values rather than begging the question of what the "right" concept is by building it into the method. Second, the method should be practical, in terms of time and effort required; for routine or everyday use by small and large services as well as for one-time studies aimed at obtaining "definitive" measurements. Third, if the measurements obtained are to serve as a basis for decisions, one should know how much confidence they merit--that is, their reliability, or reproducibility, should be statistically determininant--and this reliability should be adequate to warrant basing important decisions on the measurements. Fourth, the method should be flexible in that it can accommodate different types of indexing--for example, "keyword" indexing with no restrictions on allowable terms, subject headings controlled by an authority

* The development of this method was described in: Schultz, Claire K., Schultz, Wallace L., and Orr, Richard H., "Comparative indexing: terms supplied by biomedical authors and document titles." American Documentation 16, 4, (October 1965), pp. 299-312.

The rationale underlying the development of the method is given in Appendix A.

LIST WITH OR WITHOUT HIERARCHICAL STRUCTURE, INDEXING DONE BY PEOPLE OR BY MACHINE, ETC. FIFTH, IT SHOULD BE SENSITIVE ENOUGH TO DETECT DIFFERENCES IN THE RELATIVE MERIT OF INDEXING PRODUCED BY THE ALTERNATIVE PROCEDURES OR AGENTS THOSE USING THE METHOD MAY WISH TO ASSESS. COLLECTIVELY, THESE FIVE GENERAL DESIDERATA--FACE VALIDITY, PRACTICABILITY, RELIABILITY, FLEXIBILITY, AND SENSITIVITY--REPRESENT A STRINGENT SET OF REQUIREMENTS A TRULY GENERAL METHOD SHOULD MEET. IN ANY PARTICULAR APPLICATION, OF COURSE, THERE MUST ALWAYS BE TRADE-OFFS BETWEEN VALIDITY AND PRACTICALITY, AND BETWEEN RELIABILITY AND PRACTICALITY; HOWEVER, IT SHOULD BE POSSIBLE TO ACHIEVE COMPROMISES THAT ARE ACCEPTABLE. THIS STUDY AIMED AT EXPLORING THE METHODOLOGIC VARIABLES THAT GOVERN THE TRADE-OFFS REQUIRED AND INFLUENCE THE METHOD'S FLEXIBILITY AND SENSITIVITY.

ORGANIZATION OF REPORT

IN THE SUCEEDING SECTIONS OF THIS REPORT, WE WILL DESCRIBE THE BASIC OPERATIONS REQUIRED TO APPLY THE GENERAL METHOD; GIVE THE RESULTS OF TRIALS AND ANALYSES DESIGNED TO EXPLORE CRITICAL METHODOLOGIC VARIABLES, DISCUSS THE IMPLICATIONS OF THESE FINDINGS AS THEY RELATE TO THE DESIDERATA SET FORTH ABOVE, AND OFFER SOME CONCLUSIONS REGARDING THE METHOD'S POTENTIAL RANGE OF APPLICATIONS. FOR CLARITY OF PRESENTATION, ALL SUBSIDIARY DETAIL WILL BE RELEGATED TO THE APPENDICES.

ESSENTIALS OF METHOD

In the simplest terms, the criterion-group method can be described as follows: For each document in the test corpus a set of terms characterizing that document is first established by merging all terms chosen by the members of a criterion group, each of whom makes his choices independently. This indexing set is then considered the standard (criterion set) against which other sets of indicia (test sets) for the same document are tested. In this method the terms in the sets to be tested are not scored on a black-or-white scale--that is, they are not simply rated as "matching" or "not matching" the terms in the criterion set; our scale allows for as many shades of gray as there are members of the criterion group. Conducting a test requires six basic operations.

Selecting the document sample

In any specific application of the method, the documents for which indexing is to be evaluated should be a representative sample of the document universe of interest. This sample may be selected from this universe by any of the usual sampling procedures based on random selection. When the sample to be used is large, a simple random sampling procedure can be used; however, for small samples, a stratified random sample may be preferable. For this operation the most important variable is the size of the sample, which should be large enough to provide the reliability needed for the particular purpose. On the other hand, since the number of documents is a major factor in determining the effort and expense of running a test, this number should be no larger than necessary.

Selecting and instructing the criterion group

What type of individuals should constitute the criterion group depends upon one's concept of "ideal" or "standard" indexing and the purpose to be served. In our original study* the aim was to test how well author-supplied indicia matched the language of potential users; therefore, a group of the author's peers served as the criterion group. However, it might be considered appropriate

* Schultz, Claire K., Wallace L. Schultz, and Richard H. Orr. Comparative indexing: terms supplied by biomedical authors and document titles. American Documentation 16, 4, (October, 1965). pp. 299-312.

FOR THE CRITERION GROUP TO CONSIST OF "EXPERT" INDEXERS SELECTED ON SOME BASIS FOR THE QUALITY OF THEIR WORK. IDEALLY, FROM WHATEVER UNIVERSE THE CRITERION GROUP IS DRAWN, THE SELECTION PROCEDURE SHOULD INSURE THAT THE GROUP IS REPRESENTATIVE OF THAT UNIVERSE; BUT PRACTICAL CONSTRAINTS MAY REQUIRE ONE TO SETTLE FOR SELECTING MEMBERS OF THE GROUP BY NON-RANDOM PROCEDURES. OTHER THINGS BEING EQUAL, THE LARGER THE GROUP THE MORE LIKELY IT WILL BE REPRESENTATIVE; AND A UNIVERSE THAT IS RELATIVELY HOMOGENEOUS CAN BE ADEQUATELY REPRESENTED BY A SMALLER CRITERION GROUP THAN A UNIVERSE THAT IS HETEROGENEOUS. THE SIZE OF THE CRITERION GROUP, LIKE THE SIZE OF THE DOCUMENT SAMPLE, AFFECTS THE COST OF USING THE METHOD; THEREFORE, THIS VARIABLE IS ALSO AN IMPORTANT DETERMINANT OF PRACTICALITY.

ANOTHER VARIABLE IN THIS OPERATION IS HOW THE GROUP IS INSTRUCTED TO CARRY OUT ITS TASK, INCLUDING WHETHER THEY ARE GIVEN ANY SORT OF A TERMINOLOGY "GUIDE" EXPLICITLY OR IMPLICITLY INTENDED TO STRUCTURE THEIR RESPONSES.

INSTRUCTING TEST INDEXERS

IN ANY APPLICATION WHERE AN INDIVIDUAL, A GROUP OF INDIVIDUALS, OR A MACHINE INDEXES DOCUMENTS FOR THE SPECIFIC PURPOSE OF TESTING THE RESULTING INDICIA, INSTRUCTIONS OR RULES ON HOW TO CARRY OUT THE TASK WILL HAVE TO BE GIVEN. THESE INSTRUCTIONS MAY OR MAY NOT BE EQUIVALENT TO THOSE GIVEN THE CRITERION GROUP. IN APPLICATIONS WHERE THE INDEXING TO BE TESTED HAS BEEN PRODUCED AS PART OF AN ONGOING SERVICE, THIS VARIABLE DOES NOT REPRESENT A TEST "OPTION". AGAIN, IF ONE DESIRES TO GENERALIZE FROM THE FINDINGS REGARDING THE QUALITY OF THE TESTED INDEXING TO SOME LARGER UNIVERSE, THE QUESTION OF REPRESENTATIVENESS ARISES; THEN, THE METHOD OF SELECTION AND SIZE OF THE GROUP REQUIRE CAREFUL CONSIDERATION.

ESTABLISHING CRITERION AND TEST SETS

IF EITHER THE CRITERION GROUP OR THE TEST INDEXERS ARE ALLOWED TO USE FREE LANGUAGE*, A DECISION IS REQUIRED ON WHETHER THEIR OUTPUT SHOULD BE EDITED, OR STANDARDIZED, BEFORE CRITERION AND TEST SETS ARE COMPARED; AND IF STANDARDIZING IS DONE, WHAT RULES SHOULD BE FOLLOWED. WITHOUT STANDARDIZATION, SYNOMYMS AND TRIVIAL VARIATIONS--FOR EXAMPLE, SINGULAR AND PLURAL FORMS OF THE SAME TERM--WILL BE

* OR IF MACHINE INDEXING IS TO BE TESTED

COUNTED AS DIFFERENT TERMS. HOWEVER, ANY EDITING INCREASES THE COST OF A TEST; AND ALL HUMAN EDITING IS PRONE TO INCONSISTENCIES AND BIASES THAT MAY AFFECT THE RELIABILITY AND THE VALIDITY OF TEST RESULTS.

WEIGHTING THE CRITERION SETS

IN THIS METHOD, SOME SCHEME IS REQUIRED FOR WEIGHTING THE TERMS USED FOR INDEXING A DOCUMENT TO REFLECT THE CONSENSUS THAT EXISTS AMONG THE CRITERION GROUP WITH RESPECT TO APPROPRIATE INDEXING TERMS FOR THAT DOCUMENT. MANY SCHEMES COULD BE EMPLOYED, BUT PERHAPS THE SIMPLEST IS TO WEIGHT EACH TERM IN THE CRITERION SET BY THE NUMBER OF CRITERION GROUP MEMBERS WHO USED IT TO CHARACTERIZE THE DOCUMENT AND TO GIVE ANY TERM NOT USED BY AT LEAST ONE MEMBER OF THE CRITERION GROUP (THAT IS, ANY TERM NOT IN THE CRITERION SET) A WEIGHT OF ZERO TO INDICATE ITS "UNDESIRABILITY". ALTERNATIVE SCHEMES CAN BE DEVISED THAT WILL INCREASE OR DECREASE THE EFFECT OF CONSENSUS AND WILL CHANGE THE "PENALTY" FOR USING TERMS THAT ARE NOT IN THE CRITERION SET. (SEE APPENDIX D FOR DETAILS ON WEIGHTING AND AN EXAMPLE OF AN ALTERNATIVE SCHEME.)

SCORING THE TEST SETS

THE WEIGHTS THUS ESTABLISHED ARE EMPLOYED TO SCORE EACH TEST SET BY ADDING THE WEIGHTS FOR EACH TERM IN THE SET. THE "RAW SCORE" FOR A TEST SET IS THEN STANDARDIZED BY EXPRESSING IT AS A PERCENTAGE OF THE HIGHEST SCORE POSSIBLE FOR THAT SET, OR THE "VARIABLE SCORE", WHICH IS DETERMINED BY THE SUM OF THE WEIGHTS FOR ALL TERMS IN THE CRITERION SET. THUS IF A TEST SET SCORES 0%, IT MEANS THAT NO TERM IN THE SET WAS USED BY ANY MEMBER OF THE CRITERION GROUP; AND A SCORE OF 100% MEANS THAT THE TEST SET CONTAINS ALL THE TERMS USED BY THE CRITERION GROUP COLLECTIVELY.

WHEN THE CRITERION GROUP CONSISTS OF POTENTIAL USERS, THE PERCENT MAXIMAL SCORE IS ANALOGOUS TO CLEVERDON'S "RECALL" MEASURE; AND IF DESIRED, A SUPPLEMENTARY FIGURE OF MERIT ANALOGOUS TO HIS "PRECISION" MEASURE MAY ALSO BE CALCULATED BY TAKING INTO CONSIDERATION THE FREQUENCY WITH WHICH TERMS NOT IN THE CRITERION SETS (NON-SCORING OR "ZERO TERMS") APPEAR IN THE TEST SETS. * (SEE APPENDIX D FOR DETAILS ON SCORING.)

* THE RELATION OF MEASURES DERIVED BY THIS METHOD TO OTHER MEASURES OF INDEXING PERFORMANCE ARE SUGGESTED IN APPENDIX A. A FULL DISCUSSION OF THESE RELATIONS IS OUTSIDE THE SCOPE OF THIS REPORT.

IN THIS OPERATION, ONE MAY WISH TO GIVE SOME CREDIT FOR TEST SET TERMS THAT, ALTHOUGH NOT IDENTICAL TO TERMS IN THE CRITERION SET, ARE SUBSUMED BY CRITERION SET TERMS IN A GIVEN INDEXING VOCABULARY, THE METHOD ALLOWS THE OPTION OF DEALING WITH SUCH MISMATCHES BY "CONFOUNDING" OR "GENERIC POSTING" BEFORE SCORING THE TERM SETS. * THIS COMPLICATES SCORING AND HENCE INCREASES THE COST OF A TEST; BUT IT MAY BE APPROPRIATE IN SOME APPLICATIONS.

* ALTERNATIVELY, GENERIC-SPECIFIC TRANSFORMATIONS AS WELL AS STANDARDIZATION OF SYNOMYMS MAY BE DONE IN THE EDITING OPERATION.

FINDINGS ON METHODOLOGIC VARIABLES

THIS STUDY FOCUSED ON SIX OF THE METHODOLOGIC VARIABLES SELECTED FROM THOSE IDENTIFIED ABOVE. THESE SIX VARIABLES WERE SELECTED BECAUSE, FOR A PRIORI REASONS, WE FELT THEY COULD BE MAJOR DETERMINANTS OF THE METHOD'S PRACTICALITY, FLEXIBILITY, RELIABILITY, AND SENSITIVITY, AND BECAUSE THEY COULD BE INVESTIGATED WITHOUT ESTABLISHING A NEW DOCUMENT CORPUS. TO EXPLORE THE EFFECTS OF THESE VARIABLES, WE CARRIED OUT SPECIAL ANALYSES OF THE DATA OBTAINED IN THE ORIGINAL APPLICATION OF THE METHOD AND ALSO CONDUCTED TRIALS TO OBTAIN NEW DATA BEARING ON THESE VARIABLES. THE MAJOR FINDINGS ARE SUMMARIZED AND DISCUSSED BELOW. DETAILS ON THE MATERIALS, SUBJECTS, AND MANUAL AND COMPUTER PROCEDURES REFERRED TO ARE GIVEN IN THE APPENDICES.

VARIABLE 1 -- SIZE OF DOCUMENT SAMPLE

AN INDICATION OF HOW THE RELIABILITY OF TEST SCORES DEPENDS ON THE SIZE OF THE DOCUMENT SAMPLE USED FOR A TEST IS PROVIDED BY THE STANDARD DEVIATION OF THE % MAXIMAL SCORES FOR INDIVIDUAL DOCUMENTS FROM THE MEAN SCORE FOR ALL DOCUMENTS IN THE TEST SAMPLE. WE FOUND THAT THE SAMPLE STANDARD DEVIATION IS MODERATELY AFFECTED BY OTHER METHODOLOGIC VARIABLES. ASSESSING THE EFFECTS OF EACH VARIABLE ON RELIABILITY SINGLY AND IN COMBINATION WITH OTHER VARIABLES WAS NOT FEASIBLE; HOWEVER, THE EFFECTS OF THE TECHNIQUE USED FOR EDITING TERM SETS (VARIABLE 4) AND OF THE SCHEME EMPLOYED FOR WEIGHTING BEFORE SCORING (VARIABLE 5) WERE EXPLORED AND WILL BE DISCUSSED LATER IN CONNECTION WITH THESE VARIABLES. FOR THE STUDIES REPORTED IN THIS SECTION AND IN THE SECTION DEVOTED TO CRITERION GROUP SIZE (VARIABLE 2), THE EDITING TECHNIQUE AND WEIGHTING SCHEME REMAINED CONSTANT.*

WHEN SETS OF TERMS PRODUCED BY 8 PROFESSIONAL INDEXERS FOR EACH DOCUMENT WERE SCORED AGAINST THE SET OF TERMS SUPPLIED BY THE CRITERION GROUP OF 12 COLLECTIVELY, THE STANDARD DEVIATION OF SCORES FOR TERM SETS AVERAGED OVER THE 8 INDEXERS FROM THE GRAND MEAN FOR A SAMPLE OF 128 DOCUMENTS WAS 17 POINTS (% OF MAXIMAL SCORE). IN A SAMPLE OF 32 DOCUMENTS, THE CORRESPONDING STANDARD DEVIATIONS FOR SCORES OF TERM SETS PRODUCED BY INDIVIDUAL INDEXERS RANGED FROM 16 TO 20. FOR TERM SETS SUPPLIED BY AUTHORS, THE STANDARD DEVIATION OF INDIVIDUAL TERM SET SCORES FROM THE MEAN FOR 256 DOCUMENTS WAS 17 POINTS. FIGURE 1 GIVES THE CONFIDENCE LIMITS FOR MEAN SCORES BASED ON DIFFERENT SAMPLE SIZES WHEN THE OBSERVED SAMPLE STANDARD DEVIATION OF 17 POINTS IS TAKEN AS AN ESTIMATE OF THE STANDARD DEVIATION FOR THE DOCUMENT POPULATION FROM WHICH THE SAMPLES WERE DRAWN.

* COMPUTER EDITING AND WEIGHTING SCHEME #1 WERE EMPLOYED THROUGHOUT.

Figure 1

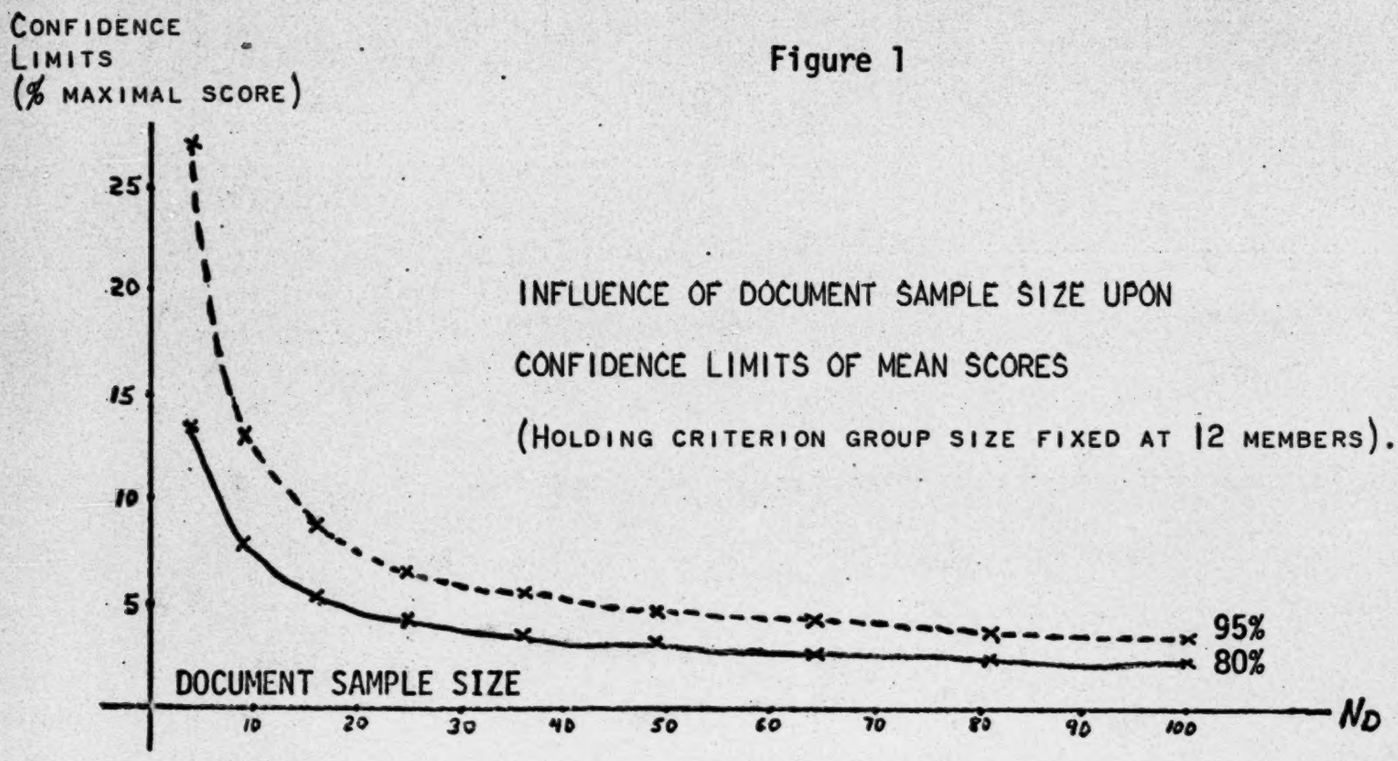
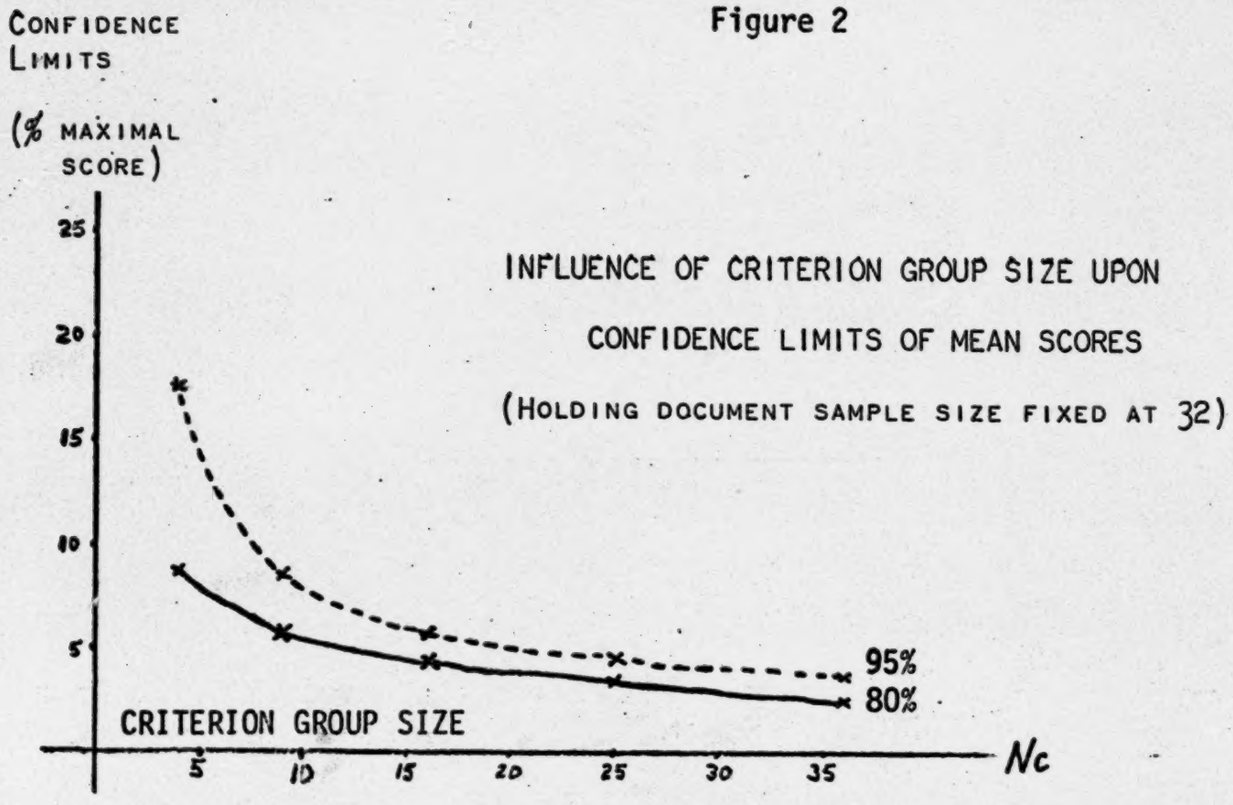


Figure 2



THE RANDOM VARIATION IN SCORES ATTRIBUTABLE TO DOCUMENTS WILL, OF COURSE, DEPEND UPON THE HETEROGENEITY OF THE DOCUMENT POPULATION FROM WHICH THE SAMPLE WAS DRAWN; AND SINCE THE PRESENT STUDIES WERE LIMITED TO OUR DOCUMENT CORPUS, ONE CANNOT SAY THE OBSERVED SAMPLE VARIATION IS A GOOD ESTIMATE OF THE VARIATION THAT WILL BE ENCOUNTERED IN APPLICATIONS OF THE METHOD WITH OTHER DOCUMENT POPULATIONS. HOWEVER, THESE FINDINGS SHOULD PROVIDE AT LEAST A ROUGH IDEA OF THE GENERAL SIZE OF DOCUMENT SAMPLE REQUIRED IN APPLICATIONS WHERE IT IS IMPORTANT TO MEASURE THE EFFECTIVENESS OF A GIVEN INDEXING "TREATMENT" WITHIN SPECIFIED CONFIDENCE LIMITS. IT CAN BE SEEN THAT WHERE THERE IS A NEED FOR RELATIVELY PRECISE MEASUREMENTS, E.G., WITHIN ±5 POINTS AT THE 95% CONFIDENCE LEVEL, SAMPLES OF 50 TO 100 DOCUMENTS WILL PROBABLY SUFFICE UNLESS VARIATION IN THE DOCUMENT POPULATION IS CONSIDERABLY GREATER THAN IN OUR CORPUS. FOR MANY APPLICATIONS, THIS DEGREE OF PRECISION WILL NOT BE NECESSARY AND USEFUL RESULTS CAN BE OBTAINED WITH CONSIDERABLY SMALLER SAMPLES,--FOR EXAMPLE, WHERE A ROUGH ESTIMATE (± 10 POINTS WITH 80% CONFIDENCE) CAN BE USEFUL, SAMPLES OF 10 DOCUMENTS MAY SUFFICE.

MANAGERS OF INDEXING SERVICES AND RESEARCHERS ATTEMPTING TO DEVELOP INDEXING "SYSTEMS" OFTEN NEED TESTS TO INDICATE WHETHER TWO INDEXING TREATMENTS GIVE SIGNIFICANTLY AND MATERIALLY DIFFERENT RESULTS. FOR SUCH USES, TESTS WITH SMALL SAMPLES SHOULD PROVIDE AN ADEQUATE BASIS FOR WORKING DECISIONS ON MATTERS WHERE THE COST OF BEING WRONG IS NOT GREAT. THE USE OF SMALL SAMPLE TESTS THAT TAKE ADVANTAGE OF THE REDUCED VARIABILITY ACHIEVED BY EMPLOYING THE SAME SAMPLE TO TEST TWO DIFFERENT TREATMENTS WILL BE ILLUSTRATED LATER.

VARIABLE 2 --SIZE OF CRITERION GROUP

ONE COULD ASSESS THE EFFECT OF THIS VARIABLE DIRECTLY BY SEEING HOW THE SCORES OF A GIVEN INDEXING TREATMENT FOR A GIVEN DOCUMENT SAMPLE CHANGE AS THE NUMBER OF INDIVIDUALS IN THE CRITERION GROUP INCREASES. HOWEVER, WHEN SCORING IS DONE MANUALLY AND THE DOCUMENT SAMPLE IS OF ANY SIZE, THE WORK REQUIRED FOR EACH INCREMENT IN THE SIZE OF THE CRITERION GROUP IMPOSES SEVERE LIMITATIONS ON THIS APPROACH. FOR THIS REASON, IN OUR ORIGINAL PROJECT, WE WERE ONLY ABLE TO ASSESS THIS VARIABLE CRUDELY BY GROUPING SCORES BASED ON HALF OF OUR CRITERION GROUP OF 12 SCIENTISTS WITH SCORES BASED ON THE OTHER HALF. WITH THE DEVELOPMENT OF A COMPUTER PROGRAM FOR SCORING, SYSTEMATIC ASSESSMENT OF THE EFFECT OF CRITERION GROUP SIZE BECAME FEASIBLE; HOWEVER, THE COST OF A DEFINITIVE STUDY WAS STILL MATERIAL SO WE CONSIDERED ALTERNATIVE APPROACHES THAT WOULD BE MORE ECONOMICAL AND ALSO BE USEFUL FOR UNFINISHED STUDY OF THE CHARACTERISTICS OF INDIVIDUAL SCIENTISTS THAT MAY INFLUENCE HOW EFFECTIVE INDEXING IS FOR THEM. ALTHOUGH IN THIS METHOD DEFINITIVE SCORING OF A TEST SET OF INDEXING TERMS IS BASED ON COMPARISONS WITH A "COMPOSITE" CRITERION SET ESTABLISHED BY MERGING THE TERMS USED BY EACH MEMBER OF THE CRITERION GROUP TO DESCRIBE A

GIVEN DOCUMENT, WE HAVE DEMONSTRATED EMPIRICALLY THAT THE SCORE BASED ON A COMPOSITE CRITERION SET CAN BE USEFULLY APPROXIMATED UNDER CERTAIN CONDITIONS BY AVERAGING SCORES FOR A TEST SET BASED ON INDIVIDUAL CRITERION SETS, CONSISTING OF THE TERMS USED BY EACH CRITERION GROUP MEMBER INDIVIDUALLY.* THIS SUGGESTED ANOTHER APPROACH TO ASSESSING THE EFFECT OF CRITERION GROUP SIZE UTILIZING ANALYSIS OF VARIANCE TECHNIQUES. DETAILS OF THESE ANALYSES WOULD BE INAPPROPRIATE HERE, BUT THE MAJOR FINDINGS RELATING TO THE EFFECT OF CRITERION GROUP SIZE WILL BE SUMMARIZED VERY BRIEFLY.

THESE ANALYSES INDICATE THAT AN APPROPRIATE MODEL FOR PRESENT PURPOSES IS ONE IN WHICH THE TOTAL VARIANCE IN SCORES IS PARTITIONED INTO 3 ADDITIVE COMPONENTS ATTRIBUTABLE TO DOCUMENT VARIANCE, CRITERION GROUP VARIANCE, AND RESIDUAL ERROR. WHEN CRITERION GROUP VARIANCE IS HELD CONSTANT, THIS MODEL GIVES THE SAME ESTIMATE FOR DOCUMENT VARIANCE AS THAT OBTAINED BY "EXPERIMENTAL" OR DIRECT, DETERMINATION OF DOCUMENT SAMPLE STANDARD DEVIATION REPORTED EARLIER. WHEN DOCUMENT VARIANCE IS HELD CONSTANT, THE MODEL GIVES AN ESTIMATE FOR CRITERION GROUP VARIANCE CENTERED AROUND 121 POINTS (STANDARD DEVIATION, 11 POINTS). THE EFFECT OF SAMPLING ERROR ATTRIBUTABLE TO THIS SOURCE ON TEST SCORE RELIABILITY IS SHOWN IN FIGURE 2 WHERE THE CONFIDENCE LIMITS ARE CALCULATED FROM THIS ESTIMATED VARIANCE.

ON A A PRIORI BASIS, ONE WOULD EXPECT CRITERION GROUP VARIANCE TO DEPEND UPON THE HETEROGENEITY OF THE POPULATION THE GROUP REPRESENTS. IT HAS NOT BEEN FEASIBLE TO TEST THIS HYPOTHESIS SYSTEMATICALLY; HOWEVER, WE HAVE SCORED AUTHOR-INDEXER TEST SETS AGAINST ANOTHER CRITERION GROUP--THE 8 PROFESSIONAL INDEXERS. RATHER SURPRISINGLY, THE SAME ESTIMATE OF CRITERION GROUP VARIANCE WAS OBTAINED. THESE INDEXERS ALSO CONSTITUTE A RELATIVELY HETEROGENEOUS GROUP IN THAT THEIR APPROACHES TO INDEXING REFLECT A VARIETY OF DIFFERENT INDEXING SERVICES.

FIGURE 2 INDICATES THAT, WHEN PRECISE ESTIMATES OF INDEXING EFFECTIVENESS ARE CRITICAL, THE CRITERION GROUP WILL PROBABLY HAVE TO BE SIZABLE IF ONE IS TO HAVE MUCH CONFIDENCE THAT THEY ARE ADEQUATELY REPRESENTATIVE OF SOME LARGER POPULATION. FOR OUR ORIGINAL APPLICATION OF THIS METHOD, IT WAS IMPORTANT TO INCLUDE

*THE CONDITIONS UNDER WHICH SCORING BASED ON INDIVIDUAL CRITERION SETS APPROXIMATES SCORING BASED ON COMPOSITE CRITERION SETS ARE COMPLEX AND HAVE NOT BEEN COMPLETELY EXPLORED; HOWEVER, NUMEROUS TRIALS HAVE SHOWN THAT, WHEN WEIGHTING SCHEME #1 IS EMPLOYED THE APPROXIMATION IS GOOD AT LEAST FOR TERM SETS SUPPLIED BY OUR ORIGINAL CRITERION GROUP OF SCIENTISTS.

ENOUGH PEOPLE IN THE CRITERION GROUP THAT WE COULD BE REASONABLY CERTAIN ANOTHER SAMPLE FROM THE USER POPULATION THEY REPRESENTED WOULD NOT GIVE MATERIALLY DIFFERENT SCORES FOR THE INDEXING TREATMENTS WE WANTED TO ASSESS. WITHOUT A GUIDE AS TO HOW MANY WOULD BE "ENOUGH", WE THEREFORE MADE THE SAMPLE AS LARGE AS WE COULD WITHIN PRACTICAL CONSTRAINTS. IN ANY APPLICATION WHERE MEMBERS OF THE CRITERION GROUP ARE SUPPOSED TO REPRESENT SOME LARGE POPULATION, HOW LARGE THE GROUP SHOULD BE IS A CRITICAL CONSIDERATION SINCE THIS VARIABLE IS A MAJOR DETERMINANT OF THE OVERALL COST OF EMPLOYING THE METHOD. FOR OTHER APPLICATIONS, HOWEVER, THE REPRESENTATIVENESS OF THE CRITERION GROUP IS IRRELEVANT--FOR EXAMPLE, WHERE ONE CAN IDENTIFY A FEW "EXPERT" INDEXERS AND CONSIDER THEIR "OUTPUT" AS A VALID STANDARD. IF A CRITERION GROUP WERE SELECTED FROM THE "BEST" INDEXERS WORKING FOR A SINGLE SERVICE, IT SEEMS REASONABLE TO PREDICT THAT THEIR VARIANCE WILL BE MATERIALLY SMALLER THAN THAT FOUND IN THE TWO GROUPS WE STUDIED AND THAT A GROUP OF 3 OR 4 WILL PROBABLY BE OPTIMAL. EVEN WHERE THE CRITERION GROUP IS SUPPOSED TO REPRESENT SOME LARGER POPULATIONS, THERE ARE NUMEROUS POTENTIAL APPLICATIONS WHERE HIGH PRECISION IS NOT ESSENTIAL AND A CRITERION GROUP OF LESS THAN 10 MEMBERS WILL PROBABLY SUFFICE--WHERE ONLY ROUGH ESTIMATES ARE REQUIRED OR THE NEED IS FOR A QUICK TEST TO GUIDE THE KIND OF WORKING DECISIONS DISCUSSED IN CONNECTION WITH DOCUMENT SAMPLE SIZE.

VARIABLE 3--INSTRUCTIONS TO TEST INDEXERS

WHETHER THE METHOD COULD ACCOMMODATE INDEXING DONE WITHOUT ANY VOCABULARY GUIDE, SUCH AS THE AUTHOR-INDEXING FORM EMPLOYED IN THE ORIGINAL APPLICATION, WAS AN IMPORTANT QUESTION CONCERNING THE METHOD'S FLEXIBILITY; AND WHETHER THE METHOD COULD DETECT DIFFERENCES IN INDEXING PRODUCED BY ASKING INDEXERS TO FOLLOW DIFFERENT RULES HAD A BEARING ON ITS SENSITIVITY. BOTH OF THESE QUESTIONS WERE EXPLORED IN NUMEROUS SMALL-SCALE EXPERIMENTS, IN WHICH DIFFERENT TYPES OF SUBJECTS--INDIVIDUALS WITH AND WITHOUT INDEXING EXPERIENCE, AND WITH AND WITHOUT BIOMEDICAL KNOWLEDGE--WERE ASKED TO INDEX DOCUMENT SAMPLES UNDER TRIAL CONDITIONS. THE KIND OF EVIDENCE THESE EXPERIMENTS PROVIDED RELATING TO THE TWO QUESTIONS CAN BE ILLUSTRATED BY THE RESULTS OF ONE SERIES OF EXPERIMENTS, WHICH IS SUMMARIZED IN TABLE 1. WITH NO GUIDE AND NO EXPLICIT RULES, THE MEAN SCORES FOR GROUP A AND GROUP B ON THE 10 DOCUMENTS IN SUBSAMPLE X (32% VS. 25%) WERE, AS ONE WOULD EXPECT, NOT SIGNIFICANTLY DIFFERENT.* THERE WERE

*THE SIGNED-RANK (WILCOXON) TEST WAS EMPLOYED TO TEST THE SIGNIFICANCE OF THE OBSERVED DIFFERENCE. HEREAFTER, ALL STATEMENTS CONCERNING THE SIGNIFICANCE OF DIFFERENCES ARE BASED ON THE SIGNED-RANK TEST IF THE SAME SUBSAMPLE OF DOCUMENTS WAS EMPLOYED FOR BOTH INDEXING "TREATMENTS," AND THE RANK TEST (VARIOUSLY CALLED WILCOXON T TEST OR THE MANN-WHITNEY U TEST) WAS USED WHEN THE DOCUMENT SUBSAMPLES DIFFERED.

Table I. Trials of Indexing Rules and Aids with Medical Student Subjects

Test Indexers	Document Subsample X — No Rules		Document Subsample Y — Rule 1		Document Subsample Z — Rule 2	
	Without Guide	With Guide	Without Guide	With Guide	Without Guide	With Guide
	% Max. Score	% Max. Score	% Max. Score	% Max. Score	% Max. Score	% Max. Score
Group A of Non-Professional Indexers (10 medical students)	32%	46%	--	36%	--	49%
Group B of Non-Professional Indexers (9 medical students)	25%	--	28%	--	41%	--

All test sets were weighted by Scheme #1. Each of the 3 document subsamples contained 10 documents.

Rule 1 "Please index this sample of documents the way you think their authors would index them."

Rule 2 "After studying the enclosed record of how authors actually indexed the last document sample you were given, please try again to apply Rule 1."

NO PROBLEMS IN STANDARDIZING AND SCORING TEST SETS UNDER SUCH UNSTRUCTURED CONDITIONS, AND THE MEAN SCORE FOR GROUP B REMAINED STABLE WHEN RULE 1 WAS IMPOSED FOR THEIR SECOND SUBSAMPLE (Y)--THIS RULE MAY BE CONSIDERED A CONTROL IN THAT IT WAS NOT EXPECTED TO MAKE A DIFFERENCE. HOWEVER, WHEN RULE 2 WAS IMPOSED FOR THEIR THIRD SUBSAMPLE, A SIGNIFICANT DIFFERENCE (99% CONFIDENCE) IN MEAN SCORES RESULTED. THE IMPLICATIONS OF TRIALS WITH GROUP A ARE LESS CLEAR CUT SINCE, FOR COMPARISONS OF INTEREST IN THE PRESENT CONTEXT, THE VARIABLES ARE CONFOUNDED. THESE TRIALS, IN CONJUNCTION WITH EVIDENCE PROVIDED BY SIMILAR EXPERIMENTS WITH OTHER TYPES OF SUBJECTS, INDICATE THAT THE METHOD IS INDEED FLEXIBLE ENOUGH TO ACCOMMODATE INDEXING DONE WITHOUT A VOCABULARY GUIDE AND THAT IS SENSITIVE ENOUGH TO DETECT THE EFFECT OF INDEXING RULES AND INDEXING AIDS. IN ADDITION, THESE ILLUSTRATE HOW SMALL DOCUMENT SAMPLES MAY SUFFICE FOR SOME APPLICATIONS.

VARIABLE 4--PROCEDURES FOR EDITING CRITERION AND TEST SETS.

THREE DIFFERENT PROCEDURES FOR EDITING WERE ASSESSED FOR THEIR GENERAL EFFECTS ON TEXT SENSITIVITY. THE PROCEDURES WERE AS FOLLOWS:

NO EDITING COMPLETELY UNEDITED TEST SETS WERE COMPARED WITH AND SCORED AGAINST THE UNEDITED TERM SETS OF THE CRITERION GROUP ON A WORD-BY-WORD BASIS. THIS MEANT THAT WHERE "NONSUBSTANTIVE" WORDS, SUCH AS "IN" AND "OF", WHICH WERE PRESENT IN AN UNEDITED CRITERION SET, MATCHED WORDS IN A TEST SET SCORING CREDIT WAS GIVEN. ON THE OTHER HAND, NO SCORING CREDIT WAS GIVEN IF A TEST SET WORD FAILED TO MATCH A "SUBSTANTIVE" WORD IN THE CRITERION SET BECAUSE OF A SLIGHT ORTHOGRAPHIC DIFFERENCE, E.G., A WORD ENDING.

COMPUTER EDITING THE COMPUTER FILE OF THESAURUS RULES EDITED BOTH THE CRITERION SETS AND TEST SETS TO ELIMINATE MOST "NONSUBSTANTIVE" WORDS AND TO STANDARDIZE WORD ENDINGS.

MANUAL EDITING A HUMAN EDITOR ATTEMPTED TO APPLY TO CRITERION AND TEST SETS THE THESAURUS RULES INCORPORATED IN THE COMPUTER EDITING PROGRAM; HOWEVER, THIS WAS DONE LARGELY BY MEMORY AND THE EDITOR UNDOUBTEDLY CONSIDERED A WIDER RANGE OF CONTEXTS THAN WAS AVAILABLE TO THE COMPUTER. FOR EXAMPLE, TERMS THAT DID NOT MATCH BECAUSE OF MISSPELLING WERE CREDITED BY THE HUMAN EDITOR.

INITIAL EXPLORATORY TRIALS WITH SAMPLES OF 8 DOCUMENTS SUGGESTED THAT AS COMPARED TO NO EDITING, BOTH COMPUTER AND MANUAL EDITING INCREASED THE METHOD'S ABILITY TO PICK UP DIFFERENCES IN SCORES GIVEN BY DIFFERENT INDEXING TREATMENTS OF THE SAME DOCUMENTS--FOR EXAMPLE, PROFESSIONAL INDEXERS VS. AUTHOR INDEXERS. HOWEVER, LATER TRIALS CONDUCTED WITH SAMPLES OF 32 DOCUMENTS INDICATED THAT, IN THIS REGARD, ANY ADVANTAGE OF THESE PROCEDURES OVER NO EDITING WAS RELATIVELY SMALL. SOME OF THE CRITICAL COMPARISONS IN THE LATER TRIALS ARE SHOWN IN TABLE 11.* THE PRINCIPAL EFFECT OF EDITING IS TO INCREASE SCORES FOR ALL INDEXING TREATMENTS AND THIS INCREASE IS SOMEWHAT MORE MARKED WITH MANUAL EDITING THAN COMPUTER EDITING; HOWEVER, THE DIFFERENCES BETWEEN MEAN SCORES FOR TWO DIFFERENT INDEXING TREATMENTS IS NOT UNIFORMLY INCREASED. IN ADDITION, THE STANDARD DEVIATIONS, WHICH ALSO AFFECT SENSITIVITY, ARE GENERALLY INCREASED BY BOTH COMPUTER AND HUMAN EDITING PROCEDURES. IT IS OF SOME INTEREST TO NOTE THAT THE COMPUTER PROGRAM QUITE SUCCESSFULLY SIMULATED A HUMAN EDITOR; THE MEAN SCORE OF ALL PROFESSIONAL INDEXER TEST SETS OVER 32 DOCUMENTS WAS 34 (STANDARD ERROR, 1.1) WHEN EDITED BY COMPUTER, AS COMPARED TO 36 (STANDARD ERROR, 1.5) WHEN THE SAME TEST SETS WERE MANUALLY EDITED.

ALTHOUGH THE FACT THAT WITHOUT EDITING NON-SUBSTANTIVE WORDS PRESENT IN THE CRITERION SET ARE COUNTED IN SCORING MAY OFFEND ONE'S INTUITIVE SENSE OF TEST VALIDITY, THE FINDINGS SEEM TO INDICATE THAT EDITING MAKES A RELATIVELY SMALL CONTRIBUTION TO TEST SENSITIVITY. BOTH HUMAN AND COMPUTER EDITING IS RELATIVELY COSTLY; THE FORMER SHOULD BE DONE BY EXPERIENCED INDEXERS, AND THE LATTER IS DEFINITELY UNECONOMIC UNLESS LARGE VOLUMES OF TEST SETS ARE TO BE PROCESSED OR A SUITABLE THESAURUS PROGRAM HAS ALREADY BEEN WRITTEN. IF EDITING IS OMITTED, THE REMAINING OPERATIONS CAN BE CARRIED OUT BY CLERICAL PERSONNEL. THIS IS A PRACTICAL CONSIDERATION THAT MAY BE IMPORTANT FOR SOME APPLICATIONS. HAVING THE OPTIONS OF NO EDITING, COMPUTER EDITING, OR HUMAN EDITING INCREASES THE METHOD'S FLEXIBILITY AND RANGE OF POTENTIAL APPLICATIONS.

* EACH OF THE CONTRASTS SHOWN WERE ALREADY KNOWN TO BE SIGNIFICANT FROM LARGE SAMPLE TESTS WITH 128 TO 282 DOCUMENTS BUT THE DIFFERENCES WERE OF AN ORDER THAT MIGHT POSE A "CHALLENGE" FOR SMALL SAMPLE TESTS.

TABLE II COMPARISONS TO ASSESS EFFECT OF EDITING PROCEDURES ON TEST SENSITIVITY

	None		Computer		Manual	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Professional Indexers vs. Author Sets						
P.I.	26	8	34	10	36	9
Author	34	11	37	12	54	16
Difference	S*		S		S*	
Author vs. Title Sets						
Author	(not done)		(not done)		54	16
Title					34	17
Difference					S*	

*S = significant at 80% level or higher; where there is an asterisk the difference was also significant at the 95% level or higher. All scores are % maximal scores (weighting Scheme I) on the same 32 documents, and the standard deviations in percentage points are given in parentheses below each score. Standard deviations were calculated as described earlier in the section devoted to the effect of document sample size.

Variable 5--Weighting scheme for scoring

Trials of two different weighting schemes were conducted primarily to determine whether test sensitivity was affected by this variable. In Scheme 1, terms in test sets are weighted by the frequency with which they were used by the members of the criterion group in describing the given document; whereas, in Scheme 2, which was the one employed in the original application, the square of this frequency is used for weighting. It can be seen that the latter scheme places much greater emphasis on "popular" criterion group responses. Like editing, Scheme 2 has the effect of raising the scores of most test sets and generally increases the standard deviations in comparison with Scheme 1; however it also commonly increases the differences between mean scores for different indexing treatments. The resulting effect on test sensitivity is complex. As one example, with Scheme 1 mean scores for computer-edited professional indexer sets vs. author sets are 34 (s.d., 10) vs. 37 (s.d., 12); whereas, with Scheme 2 the corresponding values are 49 (s.d., 16) vs. 59 (s.d., 18). For this contrast, the advantage of Scheme 2 is apparent, but not marked. On the other hand, for the constrast between manual-edited author sets vs. title sets, Scheme 2 is greatly superior--54 (s.d., 16) vs. 34 (s.d., 17) as compared to 74 (s.d., 16) vs. 35 (s.d., 26). Since weighting by Scheme 2 entails a relatively small increment in effort over what is required with Scheme 1, it may be a useful option in some circumstances.

We considered weighting schemes that would "penalize" overassignment of terms more heavily than either Scheme 1 or Scheme 2; for example, by giving a negative weight to terms in test sets that were not used by any number of the criterion group. However, the schemes considered had numerous technical disadvantages; and since the % maximal score divided by the total number of terms in the test set can serve as a measure of indexing efficiency, as contrasted with effectiveness, this matter has not been pursued further.

Variable 6--Confounding before scoring

As an exploratory trial of the effect of confounding, i.e., generic posting, all professional indexer and author sets for 32 documents were rescored after adding to each test set any terms shown by the vocabulary guide as generic to terms in the original test set. Scoring credit was then given to such added terms when they matched criterion set terms. Confounding increased the grand mean for the professional indexer sets by 6 points, and the mean for author sets was also increased by 6 points; in both cases, the standard deviation was unchanged. These findings suggest that confounding has little or no effect on test sensitivity, which was the main question prompting the trial. Confounding, however, may have an advantage for certain applications, e.g., in tests in the context of an indexing system that employs hierarchical structure, and where it may be desirable to make more comparable indexing done at different levels of specificity. If an hierarchical organization of indexing terminology has been created prior to scoring, the process can be carried out during either manual or computer scoring at a relatively low cost. Confounding therefore represents a useful option that increases the method's flexibility.

Conclusions

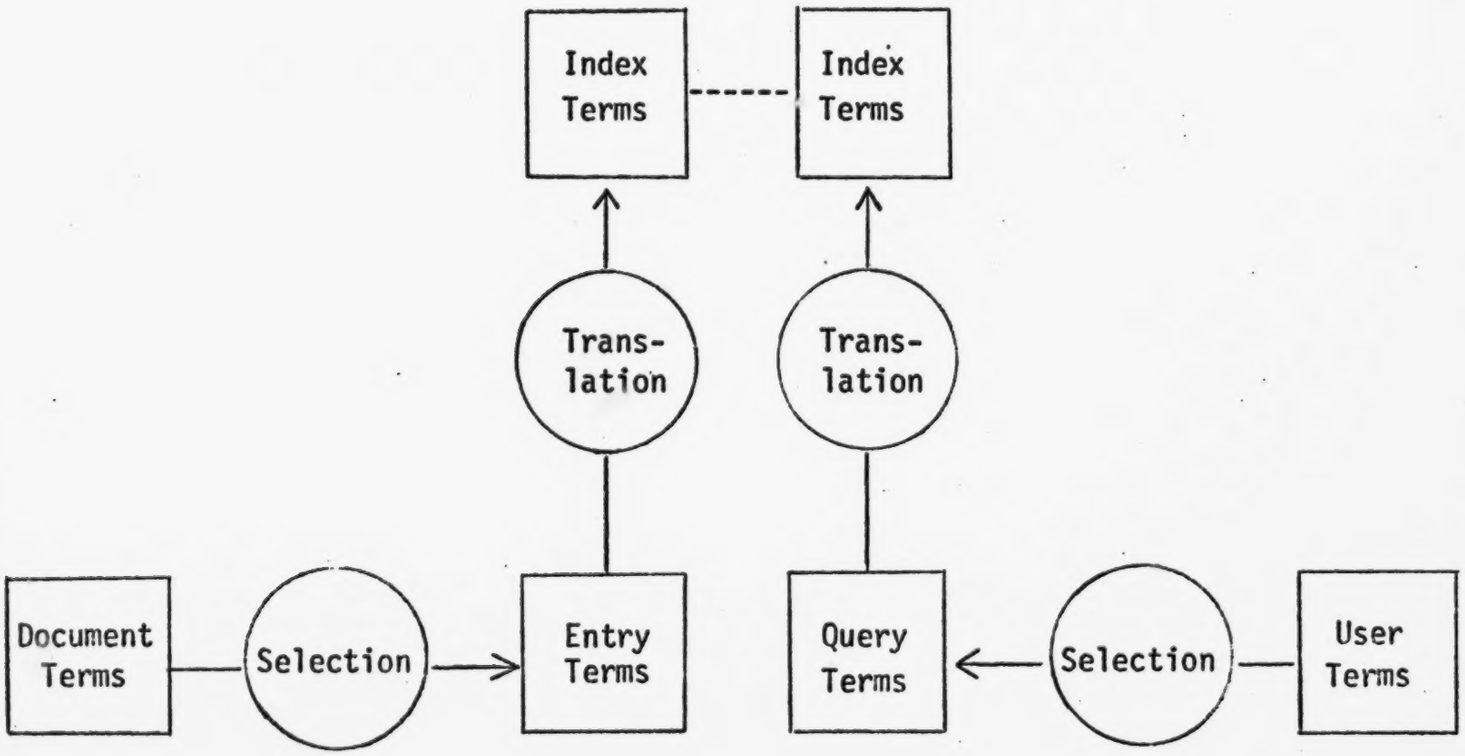
From the results of these studies of the methodologic variables, we have concluded that the consensus-group method of evaluating indexing can be a practical yardstick for a wide variety of managerial, research, and educational uses.

APPENDIX A

RATIONALE OF METHOD AND LITERATURE REVIEW

A MODEL OF INFORMATION RETRIEVAL

THE FOLLOWING DIAGRAM WHICH HAS BEEN FREELY ADAPTED FROM KYLE[18]* AND HYSLOP,[15] REPRESENTS A SIMPLIFIED# MODEL OF THE CHAIN OF PROCESSES IN AN "INFORMATION RETRIEVAL" SYSTEM. THIS MODEL CAN ACCOMODATE ANY SYSTEM IN WHICH DOCUMENTS ARE INDEXED PRIOR TO THE RECEIPT OF QUERIES, WHETHER THE INDEXING IS DONE BY PEOPLE, MACHINES, OR MAN-MACHINE COMBINATIONS.



* IN THE FOLLOWING PAGES, ONLY ONE REFERENCE IS CITED ON MOST POINTS; OFTEN SEVERAL OTHER REFERENCES WOULD BE EQUALLY APPROPRIATE. SELECTION OF THE ONE USED AS AN EXAMPLE WAS USUALLY FORTUITOUS AND IS NOT MEANT TO IMPLY ATTRIBUTION OF PRIORITY OR NOVELTY.

THE OPERATIONS OF CODING, FILING, AND MATCHING ARE OMITTED HERE; THE "SHORT-CIRCUIT" IS SYMBOLIZED BY THE DOTTED LINE BETWEEN THE TOPMOST BOXES.

In this model, indexing is depicted as a two-step process (circles labelled 1 and 2). The operations performed during the first step, selection, determine which "aspects" of the document will be represented in the index. (This step is also called "concept analysis" [9], "document analysis" [35], "detection" [3], and various other names.) The output of selection is a set of entry terms. (Synonyms for entry terms include, among others, "entry expressions" [23], "detection terms" [1], "clue words" [19], "indicator words" [12], and "candidate terms" [40].) In the second step, the entry terms are translated into a set of index terms. (This translation is commonly referred to as "standardization" or "vocabulary control".) The distinction between the two steps was aptly described by Kyle as the difference between "what to index" and "how to index it". [18]

In systems where no attempt is made to control the number of different terms that may appear in the index (for example, systems using KWIC indexing, or "pure" Uniterms [37], the translation step is, of course, missing; entry terms ARE index terms. Since few systems require indexers to record entry terms routinely, indexing may also APPEAR to be a one-step process in many systems where index terms are controlled. In such cases, however, it is reasonable to postulate that the two steps occur in the indexer's mind, even though there is evidence to suggest that professional indexers may sometimes think directly in controlled indexing language when deciding which aspects of a document should be represented in the index. [24] Despite the fact that it is often difficult to separate cleanly the selection and translation steps, the distinction is very useful in analyzing the indexing process because selection poses theoretical and practical problems of a different order of difficulty than those of translation.

The importance of the selection step

Cleverdon [8] and others [20] have postulated that, given a well-developed "indexing language," * the translation step can be reduced to a clerical or machine-like routine; whereas, the selection step is an intellectual task. The fact that translation was successfully automated in 1963, [30] and that computer programs to accomplish the translation step have since been integrated into several operating services, [22,33] as well as being demonstrated (as contrasted to simulated) in experimental trials, such as, Artandi's,[1] indicated the validity of the position that the problems of selection are of a different order than those of translation.

* As a minimum, an indexing language includes a set, or vocabulary, of entry terms; a set of index terms; and rules for translating from one set to the other. Indexing languages may have various other elements, but these three are essential.

IN ADDITION TO BEING A MORE CHALLENGINGLY DIFFICULT STEP, THE EFFECTIVENESS WITH WHICH SELECTION IS CARRIED OUT FIXES THE UPPER LIMITS ON THE PERFORMANCE OF THE ENTIRE CHAIN OF PROCESSES IN AN INFORMATION RETRIEVAL SYSTEM. AGAIN, IT WAS CLEVERDON WHO, WHETHER HE WAS THE FIRST TO DO SO OR NOT, CAN BE CREDITED FOR EMPHASIZING THIS IMPORTANT POINT AND MAKING IT CONVINCING. HE POINTED OUT THAT THE MAXIMUM PERFORMANCE ANY GIVEN SYSTEM IS CAPABLE OF, WITH REGARD TO "RECALL" AND "PRECISION" # DEPENDS UPON HOW COMPLETELY AND SPECIFICALLY ALL THE "CONCEPTS" IN THE DOCUMENTS HAVE BEEN INDEXED.[8] SINCE THE COMPLETENESS (OR "EXHAUSTIVITY") AND THE SPECIFICITY OF THE INDEX TERMS FOR A DOCUMENT CAN BE LESS, BUT NO GREATER, THAN THE COMPLETENESS AND SPECIFICITY OF THE ENTRY TERMS SELECTED FOR THAT DOCUMENT, IT FOLLOWS THAT HOW THE SELECTION STEP IS DONE DETERMINES THE HIGHEST LEVEL OF PERFORMANCE A GIVEN SYSTEM CAN PROVIDE -- VARIATIONS IN THE EFFECTIVENESS OF THE TRANSLATION STEP, OF THE INDEXING LANGUAGE ITSELF, AND OF ALL OTHER PROCESSES AND COMPONENTS IN THE SYSTEM CAN ONLY LOWER SYSTEM PERFORMANCE BELOW THIS THEORETICALLY ATTAINABLE LEVEL. IN OTHER WORDS, GOOD SELECTION IS A NECESSARY BUT NOT SUFFICIENT CONDITION FOR GOOD PERFORMANCE.

WE HAVE LEARNED HOW TO USE MACHINES IN OPERATING SYSTEMS TO EXECUTE, TIRELESSLY AND WITHOUT ERROR, THE TRANSLATIONS SPECIFIED BY AN INDEXING LANGUAGE; WE ARE BEGINNING TO LEARN HOW TO DESIGN AND USE INDEXING LANGUAGES SO THAT EITHER RECALL OR PRECISION CAN BE EMPHASIZED, DEPENDING ON WHAT THE REQUESTOR WANTS; AND MARKED PROGRESS HAS BEEN MADE IN IMPROVING CODING AND FILING, THE FINAL PROCESSES ON THE INDEXING SIDE OF THE INFORMATION RETRIEVAL "CHAIN" (SEE FIGURE, PAGE A-1). RELATIVE TO THE THEORETICAL AND PRACTICAL ADVANCES IN ALL THESE AREAS, PROGRESS APPEARS TO HAVE BEEN MUCH SLOWER IN UNDERSTANDING AND IMPROVING THE SELECTION STEP OF INDEXING. ONE CAN ARGUE THAT TODAY, AT THE PRESENT STATE-OF-THE-ART, SELECTION IS THE CRITICAL PROBLEM IN INDEXING, BOTH THEORETICALLY AND PRACTICALLY. FOR THESE REASONS WE WANTED TO DEVELOP AN EVALUATION METHOD THAT FOCUSSED SPECIFICALLY ON THE SELECTION STEP AND COULD MEASURE ITS EFFECTIVENESS INDEPENDENT OF THE TRANSLATION STEP.

$$\#\ \text{RECALL} = \frac{\text{NUMBER OF RELEVANT DOCUMENTS RETRIEVED IN RESPONSE TO A QUERY}}{\text{TOTAL NUMBER OF DOCUMENTS IN THE SYSTEM THAT ARE RELEVANT TO THE QUERY}}$$

$$\text{PRECISION} = \frac{\text{NUMBER OF RELEVANT DOCUMENTS RETRIEVED IN RESPONSE TO A QUERY}}{\text{TOTAL NUMBER OF DOCUMENTS RETRIEVED}}$$

CLEVERDON ORIGINALLY CALLED THE LATTER, "RELEVANCE RATIO" BUT LATER ACCEPTED THE SUGGESTION OF OTHERS AND CHANGED IT TO "PRECISION RATIO".[9]

Review of the criterion problem in indexing *

When we were writing up the first use of our method for evaluating indexing, that is, our criterion measure, Cleverdon's "recall" and "precision" ratios still had almost the status of an international standard. At that time we were rather apologetic about introducing a new criterion measure, particularly one that had not yet been validated against the "ultimate" criterion concept, which in its fullest explication runs something like this: Performance of a real system, in a real environment, supplying real documents, from a real collection, in response to real queries, prompted by real problems of real users--with performance rated objectively on the basis of how completely the system retrieves every document in the collection that the user judges as "relevant" to his query, and how completely it relieves the user of the chore of weeding out documents he finds irrelevant. A number of trends that began several years ago have recently accelerated, and the criterion "problem" has changed markedly since our hesitant introduction of a new measure. These trends can be summarized as follows:

(1) There is growing recognition of the need for, and legitimacy of, proximate criterion measures.

(2) In addition to good retrieval performance, as measured by recall and precision, other system desiderata are receiving more emphasis.

(3) The universal appropriateness and general utility of recall and precision as measures of the ultimate criterion concept is being questioned more frequently. [36, 38]

(4) The concept of "relevance", which is central to recall and precision measures, is undergoing a rapid and drastic metamorphosis. [13, 2, 11, 10]

(5) On the most fundamental level, the implicit assumption behind the old "ultimate" criterion concept is being challenged; [4] this assumption implies that <u>exhaustive search</u> is <u>the</u> function to be served, whereas it is only one of the several functions of IR systems (e.g., alerting, browsing, searching for "enough" documents to make a decision, etc.), and not necessarily the most important one.

* Snyder's distinctions among "criterion concept", "criterion measure", and "criterion value" [35] are useful, and will be observed in the following discussion.

SINCE THE FIRST TWO TRENDS ARE ESPECIALLY PERTINENT TO THE MEASURES OF PREFERREDNESS TO BE USED IN THE PROPOSED STUDY, THEY WARRANT SOME DISCUSSION.

ALTHOUGH PROXIMATE CRITERIA CONCEPTS HAVE LONG BEEN USED FOR DAY-TO-DAY QUALITY CONTROL (E.G., ACCURACY AS JUDGED BY INDEXING SUPERVISORS), AND AS A BASIS FOR MANAGEMENT DECISIONS (E.G., QUALITY OF INDEXING AS JUDGED BY EXPERTS IN IR, OR BY EXPERTS IN THE SUBJECT-MATTER OF THE COLLECTION), THEY WERE CONSIDERED A KIND OF SECOND-CLASS MEASUREMENT AFTER RECALL AND PRECISION GAINED WIDE ACCEPTANCE IN THE IR COMMUNITY AROUND 1963. MORE RECENTLY, HOWEVER, BASED ON CONSIDERATIONS OF EXPERIMENTAL DESIGN, SNYDER [35] HAS OFFERED CONVINCING ARGUMENTS FOR THE UTILITY AND LEGITIMACY OF PROXIMATE, OR INTERMEDIATE, CRITERION CONCEPTS AND MEASURES. HE POINTS OUT THE NEED TO STUDY SEPARATELY THE DIFFERENT COMPONENTS IN AN IR SYSTEM USING SENSITIVE MEASURES SPECIFIC FOR THE COMPONENT BEING STUDIED. APPARENTLY HE IS NOT READY TO ABANDON THE OLD ULTIMATE CRITERION CONCEPT, HOWEVER, FOR HE ADDED THE PROVISO THAT ANY PROXIMATE CRITERIA SHOULD BE VALIDATED AGAINST RETRIEVAL PERFORMANCE, PRESUMABLY MEASURED IN TERMS OF RECALL AND PRECISION. THE SAME NEEDS WERE EXPRESSED IN DIFFERENT WORDS BY A STUDY CONFERENCE SPONSORED BY NSF IN FEBRUARY 1965, WHERE THE CONSENSUS WAS THAT

> "FOR THE TIME BEING, IN VIEW OF THE PRESENT STATE OF THE ART, EFFORTS TO DEVELOP AND TEST EVALUATION METHODS AND TO CONDUCT TESTS SHOULD BE CONCENTRATED ON SELECTED FEATURES OF DOCUMENT SEARCHING SYSTEMS IN SYSTEMS CONTEXT, RATHER THAN ON TOTAL SYSTEMS". [21]

IF ONE STILL HOLDS TO THE OLD ULTIMATE CRITERION CONCEPT, FOR WHICH THE PROPER MEASURES OF SYSTEM PERFORMANCE ARE EXPRESSED SOLELY IN TERMS OF TWO OR MORE RATIOS BASED ON THE FOUR-WAY PARTITION CREATED BY THE TWO DICHOTOMIES, RETRIEVED-NOT RETRIEVED AND RELEVANT-IRRELEVANT (OR AN N-WAY PARTITION, IF RELEVANCE IS RATED ON SOME SCALE), ALL OTHER MEASURES CAN BE ONLY PROXIMATE. THE VALIDITY OF SUCH MEASURES MUST, THEREFORE, DEPEND ON THEIR HAVING SOME DEPENDABLE RELATION TO THE ULTIMATE CRITERION MEASURE. IDEALLY, THIS RELATION IS DEMONSTRATED EMPIRICALLY; BUT IN PRACTICE IT IS OFTEN ASSUMED TO EXIST, EITHER BECAUSE THERE IS CONSENSUS THAT IT "SHOULD" EXIST OR BECAUSE IT FOLLOWS FROM AN ACCEPTED THEORY. SOONER OR LATER, HOWEVER, MOST PROXIMATE CRITERION MEASURES TEND TO ACQUIRE "FACE" VALIDITY AND ACHIEVE AN INDEPENDENT "STATUS" THAT IS ACCEPTED IN ALL EXCEPT THE MOST FORMAL USAGE. ANOTHER WAY PROXIMATE MEASURES BECOME INDEPENDENT OF THEIR ORIGINAL REFERENCE STANDARD IS BY BEING INCORPORATED INTO A NEW THEORY, OR BY A REDEFINING OF CONCEPTS CENTRAL TO THE OLD ULTIMATE CRITERION CONCEPT OF MEASURE. ALL OF THESE MECHANISMS ARE APPARENTLY

AT WORK IN THE IR FIELD TODAY, AND A HOST OF CRITERION MEASURES FOR INDEXING QUALITY ARE ACQUIRING STATUS.

MOST OF THESE "NEW" * MEASURES EMPLOY A GROUP OF INDIVIDUALS WHOSE COLLECTIVE RESPONSES ESTABLISH A CRITERION STANDARD AGAINST WHICH ANY "UNKNOWN" SAMPLE OF INDEXING IS MEASURED. THESE MEASURES FALL INTO THE FOLLOWING FOUR CATEGORIES, BASED ON THE TYPES OF INDIVIDUALS THAT COMPOSE THE "CRITERION GROUP". THE FOLLOWING TABLE CLASSIFIES SOME REPRESENTATIVE EXAMPLES OF SUCH MEASURES BUT DOES NOT INCLUDE THEM ALL:

COMPOSITION OF CRITERION GROUP	SIZE OF GROUP	NAME OF CRITERION MEASURE
I. EXPERTS (AUTHORITIES, ETC.)	AS FEW AS 1	"ACCURACY" [5]
II. INDEXERS	2 OR MORE	"CONSISTENCY" [25] "PRECISION OF MEANING" [39]
III. AUTHORS	NO. REPRESENTED IN DOCUMENT CORPUS	"RELEVANCE" [14]
IV. USERS (SIMULATED QUERIES)	AS FEW AS 1	"RELEVANCE" [9] "REPRESENTATIVENESS" [16, 17, 26]

CATEGORY IV INCLUDES ALL MEASURES IN WHICH THE CRITERION GROUP MEMBERS ARE NOT ACTUAL SYSTEM USERS JUDGING THE RELEVANCE OF SYSTEM RESPONSES TO THEIR OWN QUERIES, WHICH WERE GENERATED IN THE COURSE OF THEIR REGULAR WORK. THEREFORE, IT INCLUDES MEASURES IN WHICH THE CRITERION GROUP CONSISTS OF INDIVIDUALS WHO ARE ENTITLED TO, OR MIGHT BE EXPECTED TO, USE THE GIVEN SERVICE (POTENTIAL USERS), E.G., THE QUERISTS IN CLEVERDON'S LAST STUDY,[9] OR INDIVIDUALS FROM A POPULATION CONSIDERED COMPARABLE TO THE SYSTEM'S CLIENTELE (SIMULATED USERS). THE "FACE" VALIDITY OF THESE MEASURES DEPENDS UPON ONE'S OPINION ON HOW CLOSELY THEY APPROACH REALITY.

THE METHODS USED TO CALCULATE CRITERION VALUES, AND THE SIZE OF CRITERION GROUPS, VARY WIDELY WITHIN A GIVEN CATEGORY, AS DOES THE PROBABLE RELIABILITY OF THE VALUES OBTAINED. IN A FEW CASES THESE MEASURES HAVE BEEN EMPIRICALLY VALIDATED AGAINST RETRIEVAL PERFORMANCE IN AN OPERATING SYSTEM. AS AN EXAMPLE, IN ONE SYSTEM, WHERE AN EXPERT'S JUDGEMENT COULD BE TAKEN AS FINAL AND "CORRECT", BRYANT DEMONSTRATED THAT "ACCURACY" VALUES CORRELATED HIGHLY WITH ACTUAL RETRIEVAL PERFORMANCE, AND THAT CONSISTENCY VALUES CORRELATED HIGHLY WITH ACCURACY. [5]

* SOME ARE ACTUALLY OLD; BUT THE CONDIFENCE WITH WHICH THEY ARE USED SEEMS NEW.

ANOTHER WAY TO DEVELOP WHAT MIGHT BE CONSIDERED A SPECIAL TYPE OF A PROXIMATE CRITERION MEASURE IS TO "ADAPT" THE OVER-ALL SYSTEM CRITERION MEASURE (I.E., THE ULTIMATE CRITERION MEASURE) FOR ASSESSING SEPARATELY THE PERFORMANCE OF SOME SINGLE SYSTEM COMPONENT, SUCH AS INDEXING. IN A SPECIAL TEST SYSTEM, SUCH AS CLEVERDON'S OR THE COMPARATIVE SYSTEMS LABORATORY OF WESTERN RESERVE,[27] THIS CAN BE DONE BY ACTUALLY PERFORMING ALL THE OPERATIONS CONCERNED IN INFORMATION RETRIEVAL, KEEPING EVERYTHING CONSTANT EXCEPT THE COMPONENT UNDER STUDY.* HOWEVER, THIS PROCEDURE IS VERY DIFFICULT AND EXPENSIVE; AND ONE ALTERNATIVE IS TO "SIMULATE" THE CONDITION OF "ALL OTHER THINGS BEING EQUAL" BY SIMPLY SHORT-CIRCUITING PART OF THE CHAIN. THIS IS WHAT CLEVERDON DID IN HIS LAST SERIES OF STUDIES.[9] HIS "SEARCHES" WERE PERFORMED BY PAPER-AND-PENCIL SIMULATION ON DOCUMENT-TERM MATRICES; THIS SIMULATION, ALTHOUGH IT WAS CARRIED OUT BY PEOPLE IN THIS STUDY WAS A CLERICAL OPERATION HE STATES COULD HAVE BEEN AUTOMATED. AN ESPECIALLY INTERESTING EXAMPLE OF A SHORT-CIRCUIT STRATEGY IS AN INGENIOUS METHOD KATTER HAS DEVELOPED TO ANALYZE AND COMPARE DOCUMENTS OR DOCUMENT REPRESENTATIONS, SUCH AS, INDEX TERMS, ABSTRACTS, ETC.[16]

THE FINAL STEP IN STREAMLINING THE EVALUATION OF INDEXING, OF COURSE, IS TO REPLACE THE QUALITATIVE SYSTEM MODEL WITH A MATHEMATICAL FORMULATION THAT PERMITS QUANTITATIVE PREDICTIONS OF RETRIEVAL PERFORMANCE GIVEN NUMERICAL VALUES FOR THE VARIABLES. THEN ONE CAN SIMPLY "PLUG IN" THE CRITERION VALUE FOR THE COMPONENT OR PROCESS ONE IS STUDYING AND CALCULATE THE ABSOLUTE VALUE FOR RETRIEVAL PERFORMANCE PREDICTED BY THE MATHEMATICAL MODEL OR THEORY. A NUMBER OF SUCH MODELS HAVE BEEN ADVANCED AS APPROPRIATE FOR AT LEAST PART OF A TOTAL SYSTEM (E.G., SALTON'S[26] AND BRYANT'S[6], AND SOME OF THESE HAVE BEEN TESTED WITH VARYING DEGREES OF RIGOR. THIS ELEGANT WAY OF EVALUATING INDEXING SEEMS VERY ATTRACTIVE, ONCE THE UNDERLYING THEORY HAS BEEN WELL TESTED; BUT FOR ANY OF THE CURRENT MODELS, TESTING HAS THUS FAR BEEN LIMITED AND/OR CONFINED TO SPECIAL CASES WHERE SOME OF THE VARIABLES CAN BE SAFELY IGNORED.

* SNYDER EXPRESSES THE OPINION THAT, EVEN IF ONE COULD CONTROL ALL COMPONENTS OTHER THAN THE ONE BEING STUDIED, AND THEN SEE HOW CHANGES IN THIS ONE COMPONENT AFFECT RETRIEVAL, THE USUAL CRITERION MEASURES OF OVER-ALL SYSTEM PERFORMANCE WOULD BE TOO CRUDE AND INSENSITIVE TO ANSWER SOME IMPORTANT QUESTIONS ABOUT FACTORS INFLUENCING INDIVIDUAL COMPONENTS.[35, 7] WHETHER THIS OPINION IS BASED ON THE RESULTS OF THE EARLY CRANFIELD STUDIES, WHICH SEEMED TO SHOW THAT THE OVER-ALL SYSTEM CRITERION MEASURE WAS REMARKABLY INSENSITIVE, ON THE FINDINGS OF OTHER STUDIES, OR ON THEORETICAL CONSIDERATIONS IS NOT CLEAR.

The criterion-group method

Our method tests indexing as a subsystem of the information storage and retrieval process. It employs a group to set the standard against which quality is tested, as opposed to a single individual's judgment.

This criterion group can be made up of professional indexers, of author indexers, or of document users. The choice is left to the person using the method, according to his judgment of what constitutes ideal indexing. In its original application, users constituted the criterion group. [32]

REFERENCES

1. Artandi, Susan S. Book indexing by computer. Thesis. Ann Arbor, Michigan, University Microfilms, Inc., 1963.

2. Assorio, Peter G. Classification space analysis. Boulder, Colorado, University of Colorado. October, 1964. AD 608 034.

3. Bernier, Charles L. and E. J. Crane, "Correlative indexes. VIII. Subject-indexing vs. word-indexing." Journal of Chemical Documentation, Vol. 2, No. 2, April, 1962. p.117-122.

4. Bryant, Edward C., "Progress toward evaluation of information retrieval systems". In Committee for international cooperation in information retrieval among examining patent offices. Annual Meeting, Proceedings. Washington, D.C., Spartan Books, 1965. p. 362-377.

5. Bryant, Edward C., D. W. King, and P. J. Terragno. Some technical notes on coding errors. Denver, Colorado, Westat Research Analysts, Inc., July, 1963. WRA PO 7.

6. Bryant, Edward C., Donald T. Searles, and Robert H. Shumway. Some theoretical aspects of the improvement of document screening by associative transformations. Denver, Colorado, Westat Research Analysts, Inc., November 30, 1963. AD 628 191.

7. Cleverdon, Cyril W. Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. Cranfield, England, ASLIB Cranfield Research Project, October, 1962.

8. Cleverdon, Cyril W. and J. Mills. The testing of index language devices. Aslib Proceedings, Vol. 15, No. 4, April, 1963. p. 106-130.

9. Cleverdon, Cyril W., Jack Mills, and Michael Keen. Factors determining the performance of indexing systems, Vol. 1: Design. 2 parts: test and appendices. Cranfield, England, ASLIB Cranfield Research Project, 1966.

10. Cuadra, Carlos, R. V. Katter, E. H. Holmes and E. M. Wallace. Experimental studies of relevance judgments: Final report. 3 volumes. NSFC-424. Santa Monica, Calif., Systems Development Corp. June 30, 1967. TM-3520.

11. Doyle, Lauren B. Is relevance an adequate criterion in retrieval system evaluation? in H. P. Luhn, ed., Automation and scientific communication, proceedings of the Annual Meeting of the American Documentation Institute, Chicago, Ill., October, 1963. Washington, D.C., American Documentation Institute, 1963. Part 2, p.199-200.

12. Harlow, Jacques and Paul W. Abrahams. Research in information retrieval. Final report: an investigation of the techniques and concepts of information retrieval. Paramus, New Jersey. ITT Data and Information Systems Division, 31 July, 1964. AD 461 099.

13. Hillman, Donald J. Document retrieval theory, relevance and the methodology of evaluation. Report No. 1: characterization and connectivity. Bethlehem, Pa., Lehigh University, Center for the Information Sciences. 24 May, 1966.

14. Hillman, Donald J. Mathematical theories of relevance with respect to systems of automatic and manual indexing. in H. P. Luhn, ed., Automation and scientific communication, proceedings of the American Documentation Institute Annual Meeting, Chicago, Ill., October, 1963. Washington, D.C., American Documentation Institute, 1963. Part 2, p. 323-324.

15. Hyslop, Marjorie R. The ASM information retrieval system: after Cranfield. Journal of Documentation Vol. 21, No. 1, March, 1965. p. 27-42.

16. Katter, Robert V., Emory H. Holmes, and Richard L. Weiss. Experimental investigations of a method for analyzing document representation. TM-3090/000/00. Santa Monica, California, Systems Development Corporation. 9 August, 1966.

17. Katter, Robert V. Study of document representations; multidimensional scaling of indexing terms. Final report. August 31, 1967. NSF GN-544 Santa Monica, Calif. Systems Development Corp. TM 3627.

18. Kyle, Barbara R. F. Information retrieval and subject indexing: Cranfield and after. Journal of Documentation, Vol. 20, No. 2, June, 1964. p. 55-69.

19. Maron, M. E. Automatic indexing: an experimental inquiry. Journal of the Association for Computing Machinery, Vol. 8, No. 3, July 1961. p. 404-417.

20. Montgomery, Christine and Don R. Swanson. Machinelike indexing by people. American Documentation, Vol. 13, No. 4, October, 1962. p. 359-366.

21. National Science Foundation. Summary of study conference on evaluation of document searching systems and procedures. Washington, D.C., October 2-3, 1964. Washington, D. C., National Science Foundation, 10 February, 1965.

22. Newbaker, H. R. and T. R. Savage. Selected words in full title (SWIFT): a new program for computer indexing. In H. P. Luhn, ed., Automation and scientific communication, proceedings of the Annual Meeting of the American Documentation Institute, Chicago, Ill. October, 1963. Washington, D.C., American Documentation Institute, 1963. Part 2, p. 87-88.

23. O'Connor, John J. Mechanized indexing methods and their testing. Journal of the Association of Computing Machinery 11, 4, (October 1964) pp. 437-449.

24. Oliver, Lawrence H., Claude Mitchell, Edmund W. Fitzpatrick, and Herbert Jacobson. An investigation of the basic processes involved in the manual indexing of scientific documents. Bethesda, Maryland. General Electric Corporation, Information Systems Operation. February 11, 1966.

25. Rodgers, Dorothy J. A study of inter-indexer consistency. Washington, D.C., General Electric Company, Information Systems Section, 29 September, 1961.

26. Salton, Gerard. Information storage and retrieval. Scientific Report No. ISR-13. Dept. of Computer Science, Cornell University, Ithica, N.Y. January, 1968.

27. Saracevic, Tefko and Alan M. Rees. Towards the identification and control of variables in information retrieval experimentation. Cleveland, Ohio, Western Reserve University, Center for Documentation and Communication Research, January, 1966.

28. Schultz, Claire K., Compiler and Editor, Guide to current terminology in biomedical research. Federation Proceedings Vol. 24, no. 4, July-August, 1965.

30. Schultz, Claire K., Editing author-produced indexing terms and phrases via a magnetic tape thesaurus and a computer program. In Luhn, Hans P., ed., Automation and Scientific Communication, Washington, D.C., American Documentation Institute, 1963. p. 9.

31. Schultz, Claire K. and Richard H. Orr. Evaluating indexing by reference to term-choice patterns of a criterion group. International Federation for Documentation 1965 Congress, Abstracts of papers, Washington, D.C., October 10-15, 1965. p.86.

32. Schultz, Claire K., Wallace L. Schultz, and Richard H. Orr. Comparative indexing: terms supplied by biomedical authors and document titles. American Documentation 16, 4, (October, 1965). pp. 299-312.

33. Shepherd, Clayton A. Design and implementation of the American Society for Metals Mark II documentation system. In Parameters of information science, proceedings of the Annual Meeting of the American Documentation Institute, Philadelphia, Pa., October 5-8, 1964. Washington, D.C., Spartan Books, 1964. p. 347-354.

34. Slamecka, V. and J. Jacoby. Effect of indexing aids on the reliability of indexers. Final technical note AF30 (602)-2616 Documentation Inc., Washington, D.C. June, 1963.

35. Snyder, Monroe B., Anne W. Schumacher, Steven E. Mayer and M. Dean Havron. Methodology for test and evaluation of document retrieval systems: a critical review and recommendations. McLean, Va., Human Sciences Research, Inc., January, 1966.

36. Swanson, Don R. The evidence underlying the Cranfield results. Library Quarterly, Vol. 35 No. 1, January, 1965. p. 1-20.

37. Taube, Mortimer. Installation manual for the Uniterm System of Coordinate Indexing. Arlington Hall, Virginia, Armed Services Technical Information Agency. October, 1953.

38. Taube, Mortimer. A note on the psuedo-mathematics of relevance. American Documentation, Vol. 16, No. 2, April, 1965. p. 69-72.

39. Tinker, John F. Imprecision in meaning measured by inconsistency of indexing. American Documentation, Vol. 17, No. 2, April, 1966. p. 96-102.

40. Wall, Eugene. A unified engineering vocabulary for use in information dissemination, indexing, storage and retrieval. In H. P. Luhn, ed., Automation and Scientific Communication, proceedings of the Annual Meeting of the American Documentation Institute, Chicago, Ill., October, 1963. Washington, D.C., American Documentation Institute, 1963. Part 2, p. 37-38.

APPENDIX B

MATERIALS EMPLOYED IN STUDY TRIALS

Documents

The documents employed in all trials came from the same corpus, which consisted of 285 brief preliminary reports of research. The source of these documents and the selection of the corpus are described in the following paragraphs, which are adapted from the original report on the criterion group method.

Each year several thousand 10-minute oral papers reporting current biomedical research are given at the annual meeting of the Federation of American Societies for Experimental Biology. This national convention is the largest meeting for biomedical scientists, and the work presented is an excellent cross-section of U.S. biomedical research. The Federation consists of 6 societies, each representing a major, basic biomedical discipline--biochemistry, immunology, nutrition, pathology, pharmacology, and physiology. Only members of these societies may present unsolicited papers at the Federation Meeting; and the speaker must submit to his society a short summary (225 words or less) of what he plans to say. These summaries are published in a special issue of Federation Proceedings that appears just before the annual convention. Although the document submitted is called an "abstract", the term is a misnomer in that the document is not usually produced by abstracting some preexisting document. Authors most commonly prepare the summary before they have written the full text of their oral presentation. When published, such anticipatory abstracts, therefore, represent primary documents--condensed, preliminary reports that may or may not be followed at some later time by the publication of a more detailed report. The corpus for this study was selected by taking every 10th document published in the 1962 meeting issue of Federation Proceedings, volume 21, no. 2, March-April (2,854 documents in all). In this systematic sample, each of the 6 societies is represented by 9-11% of all the documents submitted to it.

Author-indexing form

The authors of papers given at this meeting were required to complete an "author-indexing form" like that illustrated in Figure 1. This is the form referred to as a vocabulary guide throughout this report.

FIGURE 1

"AUTHOR INDEXING FORM"

(THE FORM CONSISTED OF 2 PAGES; HERE THE LOWER PART OF THE FIRST PAGE AND THE TOP OF THE SECOND PAGE HAVE BEEN OMMITTED).*

Please study the subject-category list before marking. The list will be used primarily for the arrangement of the abstracts and for the production of the subject index to the abstracts. A secondary use will be for aid in programming.

Place the number "1" in the box at the left of the most specific category which classifies the area of your paper; the number "2" in the box at the left of the next most specific category. Do not mark more than two categories.

In the blanks at the end of the subject-category list, please supply four or more additional descriptive terms (words or short phrases) which can be used, besides the subject categories already selected, for further classifying and indexing the content of your paper. The terms you supply should preferably be nouns. Generic names of chemical compounds and drugs should be used, rather than trade names or jargon.

SUBJECT CATEGORIES

☐ 001 Amino Acids
☐ 002 Metabolism
☐ 003 Nutrition
☐ 004 Synthesis
☐ 005 Antigen-Antibody Reactions
☐ 006 Cross Reactions
☐ 007 Haptens
☐ 008 Immunofluorescence
☐ 009 In Vivo Reactions
☐ 010 *Cellular*
☐ 011 *Pathogenetic*
☐ 012 Non-specific Factors
☐ 013 *Complement*
☐ 014 *Properdin*
☐ 015 Precipitation
☐ 016 *Diffusion*
☐ 017 *Immunoelectrophoresis*
☐ 018 *Quantitation*
☐ 019 Antigens; Antibodies
☐ 020 Antibody Formation
☐ 021 Determinants
☐ 022 Microorganisms
☐ 023 *Bacteria*
☐ 024 *Rickettsia*
☐ 025 Polysaccharides
☐ 026 Proteins
☐ 027 Toxins
☐ 028 Transplantation
☐ 029 *Autoantibodies*
☐ 030 *Tissue antibodies*
☐ 031 Biological Oxidations
☐ 032 Cytochromes
☐ 033 Electron Transport
☐ 040 Coagulation
☐ 041 *Agents; factors*
☐ 042 *Fibrinolysis*
☐ 043 *Platelets*
☐ 044 Erythrocytes
☐ 045 *Destruction*
☐ 046 *Metabolism*
☐ 047 Groups
☐ 048 Hematopoiesis
☐ 049 Hemoglobin
☐ 050 Leukocytes
☐ 051 *Leukemia*
☐ 052 Plasma Proteins
☐ 053 *Albumin*
☐ 054 *Globulins*
☐ 055 Storage
☐ 056 Body Water
☐ 057 Bone
☐ 058 Carbohydrates
☐ 059 Chemistry
☐ 060 Metabolism
☐ 061 *Citric acid cycle*
☐ 062 *Glycolysis*
☐ 063 *Hexose phosphate path*
☐ 064 *Monosaccharide conversions*
☐ 065 *Polysaccharides*
☐ 066 *Small cycles*
☐ 067 Photosynthesis
☐ 068 Cardiovascular System
☐ 069 Atherosclerosis
☐ 070 *Experimental*
☐ 071 *Nutritional*
☐ 072 *Pathophysiology*
☐ 073 Blood Flow
☐ 074 *Cerebral*
☐ 075 *Coronary*
☐ 084 *Shock*
☐ 085 Blood Vessels
☐ 086 *Capillary exchange*
☐ 087 *Venous return*
☐ 088 *Wave transmission*
☐ 089 Blood Volume
☐ 090 *Hemorrhage*
☐ 091 *Transfusion*
☐ 092 Cardiac Drugs
☐ 093 Cardiac Muscle
☐ 094 *Disorders*
☐ 095 *Electrocardiography*
☐ 096 Cardiac Output
☐ 097 *Control*
☐ 098 *Measurement*
☐ 099 CV Disease
☐ 100 Edema
☐ 101 Lymph
☐ 102 Cell Structure; Function
☐ 103 Active Transport
☐ 104 Cell Membranes
☐ 105 Cytoplasm
☐ 106 Microsomes
☐ 107 Mitochondria
☐ 108 Nuclei
☐ 109 Cell, Tissue Culture
☐ 110 Cell Antigens
☐ 111 Metabolism
☐ 112 Neoplasms
☐ 113 Nucleic Acids
☐ 114 Chemotherapy
☐ 115 Bacterial
☐ 116 Cancer
☐ 117 Parasitologic
☐ 118 Connective Tissue
☐ 119 Disorders
☐ 127 Site
☐ 128 Drug Metabolism
☐ 129 Endocrines
☐ 130 Adrenal Cortex
☐ 131 Adrenal Medulla
☐ 132 Anterior Pituitary
☐ 133 *ACTH*
☐ 134 *Control of secretion*
☐ 135 *Gonadotropin*
☐ 136 *Somatotropin*
☐ 137 *TSH*
☐ 138 Brain Hormones
☐ 139 Glucagon
☐ 140 Insulin
☐ 141 *Diabetes mellitus*
☐ 142 *Mode of action*
☐ 143 Parathyroid
☐ 144 Posterior Pituitary
☐ 145 *Diabetes insipidus*
☐ 146 Sex Hormones
☐ 147 *Androgens*
☐ 148 *Estrogens*
☐ 149 *Progestogens*
☐ 150 Thyroid
☐ 151 *Iodine metabolism*
☐ 152 *Regulation*
☐ 153 *Thyroxine*
☐ 154 Energy Metabolism
☐ 155 Environment
☐ 156 Adaptation
☐ 157 Air Pollution
☐ 158 Altitude
☐ 159 Hibernation
☐ 160 Hyperthermia; Heat

☐ 210 *Synthesis*
☐ 211 *Transport*
☐ 212 Phospholipids
☐ 213 *Metabolism*
☐ 214 *Synthesis*
☐ 215 Sterols
☐ 216 *Metabolism*
☐ 217 *Synthesis*
☐ 262 Neurochemistry
☐ 263 Pain
☐ 264 Peripheral Nerves
☐ 265 Reflexes
☐ 266 *Axon*
☐ 267 *Conditioned*
☐ 268 Spinal Cord
☐ 269 Nitrogen Metabolism
☐ 320 Control
☐ 321 Disorders
☐ 322 Diuresis; Diuretics
☐ 323 Electrolyte Excretion
☐ 324 Glomerular Filtration
☐ 366 *Hepatitis*
☐ 367 Vitamins
☐ 368 B
☐ 369 B_{12}
☐ 370 C
☐ 371 Fat-Soluble
☐ 372 Folic Acid
☐ 373 Unidentified

ADDITIONAL DESCRIPTIVE TERMS

....

....

....

* THIS FIGURE IS REPRODUCED FROM: SCHULTZ, CLAIRE K., WALLACE L. SCHULTZ, AND RICHARD H. ORR. COMPARATIVE INDEXING: TERMS SUPPLIED BY BIOMEDICAL AUTH AND DOCUMENT TITLES. AMERICAN DOCUMENTATION 16, 4, (OCT. 1965). P:299-312.

Thesaurus

The thesaurus employed when criterion and test sets were manually standardized has been published as the "Guide to Current Terminology in Biomedical Research",* which represents an indexing vocabulary of the Federation of American Societies for Experimental Biology. The Guide lists 1,516 different terms consisting of one or more words, and specifies their closest equivalent in the indexing "languages" used by the National Library of Medicine, Defense Documentation Center, and the Division of Research Grants of National Institutes of Health. There is a high degree of "compatibility" among the indexing vocabularies of FASEB, NLM, DDC, and NIH; three-quarters of all the FASEB terms are readily translatable into both NLM and NIH languages.

* Schultz, Claire K., Compiler and Editor, Guide to current terminology in biomedical research. Federation Proceedings Vol. 24, no. 4, July-August, 1965.

APPENDIX C

SUBJECTS PARTICIPATING IN STUDY TRIALS

Criterion group

The criterion sets employed in this study were established by a group of potential users of the indexing for the document corpus -- in this case, members of the professional association from which the documents in this corpus were obtained, the Federation of American Societies of Experimental Biology, as described in Appendix A. The criterion group was a sample of the membership selected to represent the document authors' peers. Two active research workers from each of the six disciplines in the Federation were selected by Dr. Milton Lee, Executive Officer of the Federation, on the basis of their recognized standing in the research community and on the likelihood that they would be willing to participate in the study. To facilitate holding a meeting at Federation Headquarters in Bethesda at which the study could be explained and uniform instructions could be given, the original selection was limited to scientists in the Bethesda area (National Institutes of Health and Naval Medical Research Institute). One of the 12 scientists originally selected had to withdraw; he was replaced by a research worker in a pharmaceutical company, who was also well known in his discipline.

Author-indexers

As described in Appendix A, each of the authors of the documents in the corpus had supplied indexing terms when he submitted his paper. The author sets consisted of these indexing terms. The titles these authors had given the documents supplied the title sets.

Professional indexers

This group consisted of eight professional indexers, all of whom were experienced in working with biomedical documents. With one exception, they were senior personnel from indexing services or from information service departments of pharmaceutical companies. The exception was an indexer who was currently working directly with biomedical scientists in a university setting. This group represents a sample of the universe of such indexers selected largely on the basis of friendship with the present investigators.

Non-professional indexers

The non-professional indexers employed in this study consisted of two groups of second-year medical students from the same school, none of which had any experience in indexing. Ten students comprised Group A; there were nine students in Group B.

APPENDIX D

Procedures for Manual Implementation*

Editing and recording term uses

A term-use matrix, similar to that shown in Table I, was created for each document however, only two test sets - the author and title sets are shown here. In the present study, term usage by the professional indexer group and the non-professional indexer group were similarly recorded. Each term use by each of the 12 members of the criterion group and by the author was indicated by an X; the "presence" of a term in the title was similarly recorded. The two members of each of the six discipilinary pairs making up the criterion group were designated A and B.

Weighting

The criterion set of terms for a document consists of all the different terms used by members of the criterion group to describe that document; thus 285 criterion sets were established, one for each document. For the document illustrated in Table I, the criterion set contained 13 terms. Each term in a criterion set was assigned a weighting factor by one of two schemes -- in Scheme #1 the weight was equal to the number of criterion group members who had used it to describe the given document; whereas in Scheme #2, this number was squared. This weighting procedure is illustrated in Table I. Note that the weighting of a term was not affected by whether the author had or had not included it among the indicia he supplied, or by whether it was supplied by the document title. Terms in test sets that had not been used by at least one member of the criterion group we will refer to as "zero terms", since they were given a weight of 0.

Scoring

The raw score for each test set was calculated by adding the weights for all terms in the set. For example, for the author set shown in Table I, the raw score is 20 when the terms are weighted by Scheme #1, and 84 by Scheme #2. Since the number of terms, and the weighting of these terms, varies from one criterion set to another, the constraints on the raw scores also vary. To facilitate comparisons, we converted the raw scores into percentages of the highest score that could be awarded ("maximal score"), i.e., the sum of the weights for all terms in the criterion set. For the document illustrated in Table I, if the author set had contained all of the 13 terms in the criterion set, its raw score would have equaled the maximal score for this document (28 by Scheme #1, 96 by Scheme #2) the actual raw score was 71% of the maximal score when Scheme #1 was employed, and 88% with Scheme #2.

* This material is adapted from the published description of the first application of the method [American Documentation 16, 4, (October, 1965)].

TABLE D-1 Illustrative Term-Use Matrix for One Document

Term[a]	Criterion Group: Biochemist A	Biochemist B	Physiologist A	Physiologist B	Pharmacologist A	Pharmacologist B[b]	Pathologist A	Pathologist B	Immunologist A	Immunologist B	Nutritionist A	Nutritionist B	Term Weight # 1	Term Weight # 2	Test Sets: Author	Test Sets: Title
Chemotherapy—Cancer (116)		x	x		x		x	x		x	x		7	49	x	
Neoplasms				x	x			x				x	4	16	x	x
Mast cells				x					x	x			3	9	x	x
Cell, tissue culture—Neoplasms (112)	x				x								2	4	x	
Chemotherapy (114)				x			x				x		2	4		
Benadryl				x						x			2	4		
Histamine-antihistaminics (195)				x					x				2	4		
Bacteria				x									1	1		
Cell structure; function (102)	x												1	1		
Cell, tissue culture (109)	x												1	1		
Drug action (123)													1	1	x	
Hosts													0	0		x
Metabolism													0	0		x
Peritoneal fluid																
Mice																

[a] The parenthetical numbers following certain terms refer to the code numbers of the corresponding subject categories listed on the author-indexing form used by both authors and criterion groups; in these cases, the term given here may represent an "expansion" of the term that appeared on the form (see text). Terms not followed by a number are "write-ins," i.e., words or phrases not included among the subject categories listed on the printed form.

[b] This member of the criterion group did not respond to document No. 100.

APPENDIX E

COMPUTER IMPLEMENTATION

AFTER TEST INDICIA HAVE BEEN OBTAINED IT IS POSSIBLE TO ACCOMPLISH ANY OR ALL OF THE ADDITIONAL PROCESSING STEPS BY MEANS OF COMPUTER PROGRAMS, SUCH AS THOSE CONSTRUCTED FOR CARRYING OUT THIS STUDY.* FOR MACHINE PROCESSING THE FIRST REQUIREMENT IS THAT THE DATA BE MADE MACHINE-READABLE, THAT IS, KEYPUNCHED. THE INPUT FORMAT USED FOR THIS STUDY CONSISTED OF A DOCUMENT IDENTIFICATION, AN INDEXER IDENTIFICATION AND ONE INDEXING "TERM"--AN ALPHABETIC OR NUMBERIC EXPRESSION--OF ANY LENGTH, UP TO THE CAPACITY OF A PUNCHED CARD. THE SPECIFIC KEYPUNCHING INSTRUCTIONS USED ARE GIVEN IN TABLE E-1.

IF MACHINE-EDITING IS TO BE DONE, OR EVEN IF IT IS NOT, THE NEXT PROCESSING STEP IS TO "TAG" THE KEYPUNCHED INDICIA SO THAT EVERY "WORD" CAN BE IDENTIFIED WITH ITS DOCUMENT, INDEXER, AND POSITION WITHIN THE INDEXER'S TOTAL RESPONSE TO THE DOCUMENT. THE "TAGGED" UNITS (WORDS) ARE SORTED SO THEY WILL BE PROPERLY ORGANIZED FOR MATCHING EITHER THE THESAURUS, IF MACHINE STANDARDIZING IS DONE, OR IF EDITING IS TO BE OMITTED, THE CRITERION SET.#

STANDARDIZING INDICIA CAN ACCOMPLISH ANY OF THE FOLLOWING: (1) ELIMINATE WHAT ARE CONSIDERED "NONSUBSTANTIVE" WORDS, SUCH AS CONNECTIVES OR (2) CHANGE WORD VARIANTS SUCH AS SINGULAR AND PLURAL FORMS, INTO A "STANDARD" FORM, OR (3) CONVERT WHAT ARE CONSIDERED SYNONYMOUS EXPRESSIONS INTO A SINGLE "STANDARD" EXPRESSION, OR (4) ADD ADDITIONAL, POSSIBLY MORE GENERIC, WORDS TO INDICIA, SUCH AS "CARBOHYDRATES" IN RESPONSE TO THE TERM "GLUCOSE". TO PERFORM SUCH TRANSFORMATION ON THE RAW INDICIA THERE MUST FIRST BE A SET OF "REWRITE" RULES FOR ALL ANTICIPATED ENTRY TERMS $ AND ALSO A PROGRAM WHICH MATCHES

* PROGRAMS WRITTEN IN FORTRAN IV AND PL/I FOR USE ON THE IBM 360/67 COMPUTER. THE INVESTIGATORS CAN MAKE THESE PROGRAMS AVAILABLE TO INTERESTED PERSONS.

THE CRITERION DATA WILL HAVE BEEN GIVEN THE SAME TREATMENT AS THE INDICIA, PRIOR TO ANY MATCHING OF THE CRITERION AND INDEXING SETS.

$ INCLUDING THE TRIVIAL REWRITE RULE THAT RETAINS SOME TERMS IN THE SAME FORM AS WHEN ENCOUNTERED IN THE INPUT. WORDS NOT OF INTEREST CAN BE OMITTED FROM THE THESAURUS AND AUTOMATICALLY DELETED FOR REASON OF NON-MATCH, BUT SINCE THIS PRACTICE DELETES "NEW" WORDS OF INTEREST, IT IS A BETTER PRACTICE TO ACTIVELY DELETE UNWANTED TERMS AND "SAVE" NONMATCHING WORDS ON A SEPARATE LIST THAT IS PUT OUT FOR HUMAN REACTION. WITH THIS APPROACH IT IS ALSO POSSIBLE TO DECLINE TO MAKE THESAURAL REWRITE RULES FOR CERTAIN SEMANTICALLY AMBIGUOUS TERMS AND WAIT UNTIL THEIR FULL CONTEXTS ARE KNOWN (AT PROCESSING TIME) TO INSTRUCT THE COMPUTER HOW TO DEAL WITH SPECIFIC OCCURRENCES.

ENTRY TERMS WITH THE THESAURUS, AND THEN CARRIES OUT THE THESAURUS REWRITE RULES AS THEY ARE ENCOUNTERED. THE THESAURUS USED IN THIS STUDY EXAMINES ONE INPUT WORD AT A TIME, BUT IT ALSO EXAMINES CONTEXTS, SO THAT EXPRESSIONS SUCH AS "AMINO ACIDS" OR "CITRIC ACID CYCLE" CAN BE RETAINED AS UNITS. MORE INFORMATION IS GIVEN ABOUT THE RECENTLY DEVELOPED CONTEXT-DEPENDENT STANDARDIZING TECHNIQUE USED IN THIS STUDY IN A SEPARATE PAPER. *

THE FOLLOWING EXAMPLE WILL SERVE TO ILLUSTRATE AN EXAMPLE OF STANDARDIZATION. THE FOLLOWING PHRASE, "AMINO ACIDS IN RUMINANT NUTRITION" WOULD BE EXAMINED ONE WORD AT A TIME, AS EACH WAS ENCOUNTERED IN ITS ALPHABETIC ORDER WHEN PROCESSING A LIST OF ALL INDIVIDUAL WORDS MAKING UP THE INDICIA OF A TEST CORPUS. "ACIDS", ENCOUNTERED FIRST, WOULD BE HELD FOR POTENTIAL "PARTNER WORDS" TO BE ENCOUNTERED IN THE LATER PORTION OF THE ALPHABETIC LIST. "AMINO", ENCOUNTERED NEXT, WOULD, BY MEANS OF ITS SEQUENTIAL "TAG" BE IDENTIFIED AS A "PARTNER" OF ACIDS AND THE REWRITE RULE WOULD CAUSE "AMINO ACIDS" TO BECOME THE STANDARDIZED INDEX TERM. "IN" WOULD BE DELETED FROM THE LIST AS SOON AS ENCOUNTERED; "NUTRITION" WOULD BE HELD FOR POTENTIAL MATCH WITH "PARTNER WORDS". WHEN "RUMINANT" WAS PROCESSED IT WOULD BE IDENTIFIED AS A PARTNER WORD OF "NUTRITION" AND THE REWRITE RULE WOULD TRANSFORM "RUMINANT NUTRITION" INTO THE STANDARDIZED TERM "ANIMAL NUTRITION".

IF STANDARDIZATION IS NOT DONE, NON-SUBSTANTIVE WORDS SUCH AS "AND", WORD VARIANTS SUCH AS "ENZYME", "ENZYMES", "ENZYMAL", "ENZYMATIC", AND SYNONYMS SUCH AS "HEART" AND "CARDIAC" ARE CANDIDATES FOR MATCH. EVERY SUCH WORD IS TREATED AS UNRELATED TO THE OTHER WHEN THE CRITERION AND INDEXING SETS ARE MATCHED FOR SCORING. AS A RESULT, IF THE CRITERION SET CONTAINS, FOR EXAMPLE, "CARDIAC ARREST" AND THE INDEXING SET CONTAINS "HEART ARREST", THE INDEX SET WILL NOT GAIN ANY SCORE, BECAUSE OF MISMATCH - BUT THE SAME INDEXING SET COULD GAIN SCORE BECAUSE A TRIVIAL WORD SUCH AS "OF" DID MATCH IN THE TWO SETS.

THE PURPOSE OF THE SCORING PROGRAM IS TO PERFORM THE MATCHING OPERATION, FOR ONE DOCUMENT AT A TIME, BETWEEN THE INDEXING SET(S) AND THE CRITERION SET. THE PROGRAM CAN HAVE VARIOUS OPTIONS, AS IS TRUE FOR THE SCORING PROGRAM USED IN THIS STUDY, WHICH INSTRUCT THE PROGRAM TO CALCULATE SCORES FOR A SINGLE TEST SET OR GROUP OF SETS, FOR SINGLE DOCUMENTS OR FOR GROUPS OF DOCUMENTS. ANOTHER KIND OF OPTION INSTRUCTS THE PROGRAM TO CALCULATE STANDARD DEVIATIONS, MAKE "T" TESTS, OR PERFORM OTHER STATISTICAL COMPUTATIONS, AS THE REQUIRED DATA BECOME AVAILABLE DURING PROCESSING. THE PROGRAM CAN BE USED TO PRINTOUT DETAIL ABOUT THE MATCHING PROCESSES

* GOPNIK, MYRNA AND CLAIRE K. SCHULTZ. METHODS FOR THESAURUS PROCESSING OF CONTEXT-DEPENDENT SEGMENTS IN LANGUAGE. SUBMITTED FOR PUBLICATION.

AND INTERMEDIATE CALCULATIONS IT PERFORMS, OR ONLY SPECIFIED RESULTS, SUCH AS PERCENT OF MAXIMAL SCORE OR POINTS PER TERM.

IF ALL OF THE COMPUTER PROGRAMS JUST DESCRIBED ARE CONSIDERED AS A SYSTEM, WITH THE STANDARDIZATION PROCEDURE OPTIONAL, IT CAN BE SEEN THAT KEYPUNCHED RAW INDICIA CAN BE FED INTO THE COMPUTER, AND THE SCORED RESULTS OBTAINED, WITHOUT ANY MANUAL PROCESSING REQUIRED.

TABLE E-1
KEYPUNCHING INSTRUCTIONS FOR INDEXING DATA

WRITE THE DOCUMENT NUMBER IN THE FIRST THREE COLUMNS. IT IS ALWAYS A THREE DIGIT NUMBER (IT WILL RANGE FROM 001 TO 285) AND IS FOUND IN THE TOP RIGHT HAND CORNER OF THE INDEXING SHEET. DO NOT SKIP A SPACE. IN THE NEXT TWO COLUMNS WRITE THE NUMBER CODE OF THE PARTICULAR INDEXER. THIS WILL ALWAYS BE A TWO DIGIT NUMBER (IT WILL RANGE FROM 00 TO 99), AND IS FOUND IN THE TOP LEFT HAND CORNER OF THE INDEXING SHEET. DO NOT SKIP A SPACE. IN THE NEXT COLUMN WRITE THE NUMBER OF THE DISCIPLINE OF THE AUTHOR OF THE DOCUMENT. THE DISCIPLINE OF THE AUTHOR IS GIVEN ON THE EXTREME RIGHT OF THE FIRST LINE OF THE DOCUMENT IN THE FASEB BOOK. USE THE FOLLOWING CODE TO NUMBER THE DISCIPLINE:

1. PHYSIOLOGY
2. BIOCHEMISTRY
3. PHARMACOLOGY
4. IMMUNOLOGY
5. NUTRITION
6. PATHOLOGY

DO NOT SKIP A SPACE. IN THE NEXT COLUMNS ENTER ONE INDEXING TERM USED BY THAT INDEXER FOR THAT DOCUMENT. THIS WILL BE EITHER A THREE DIGIT NUMBER WHICH HAS BEEN CHECKED BY THE INDEXER ON THE SHEET OR A TERM WRITTEN IN BY THE INDEXER ON THE SPACE PROVIDED AT THE END OF THE SHEET. IF THE INDEXING TERM IS A NUMBER THE FINISHED CARD WILL CONTAIN NINE DIGITS WITH NO SPACES BETWEEN THEM. IF THE INDEXING TERM IS WRITTEN, THEN YOU MAY USE AS MUCH OF THE CARD AS NECESSARY TO RECORD IT. IF THE TERM IS FROM THOSE WRITTEN IN AND YOU HAVE TROUBLE READING THE HANDWRITING OR ARE NOT SURE OF THE SPELLING, READ THE DOCUMENT AND MOST OFTEN THE TERM IN QUESTION WILL APPEAR THERE. IF YOU CANNOT FIND THE TERM IN THE DOCUMENT, AND CANNOT DECIPHER THE HANDWRITING, KEYPUNCH YOUR BEST GUESS AND SET THE CARD ASIDE TO BE CHECKED. IN SOME CASES A WRITTEN-IN TERM WILL CONSIST OF MORE THAN ONE WORD. IF THIS IS THE CASE, LEAVE ONE SPACE BETWEEN WORDS. DO NOT INCLUDE ANY PUNCTUATION, E.G., COMMAS, PARENTHESES. DO INCLUDE HYPHENS. IF THE TERM IS A CHEMICAL FORMULA WITH SUBSCRIPTS, THEN COPY IT AS IF IT WERE ALL ON ONE LINE, E.G., CO_2 BECOMES CO2. REPEAT THE SIX DIGIT DOCUMENT-INDEXER-AUTHOR NUMBER AT THE BEGINNING OF THE NEXT CARD AND, FOLLOWING THE ABOVE FORMAT, RECORD THE NEXT INDEX TERM. GREEK LETTERS SUCH AS α, CAN BE PUNCHED A FOLLOWED BY A HYPHEN. BETA CAN BE PUNCHED B FOLLOWED BY A HYPHEN. DELTA CAN BE PUNCHED AS A D FOLLOWED BY A HYPHEN. GAMMA IS WRITTEN OUT (GAMMA) FOLLOWED BY A HYPHEN.

IF A TERM WILL EXCEED COLUMN 78 LOOK FOR A CONNECTIVE EARLIER IN THE TERM WHERE IT COULD BE BROKEN INTO TWO TERMS.

EX: AMINO ACID METABOLISM NUTRITION/OF IMMATURE EMBRYONIC CHICKS FED METHIONINE

OE 6000 (REV. 9-66)

DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
OFFICE OF EDUCATION

ERIC REPORT RESUME

(TOP)

ERIC ACCESSION NO.	
CLEARINGHOUSE ACCESSION NUMBER	
RESUME DATE	— —
P.A.	
T.A.	
IS DOCUMENT COPYRIGHTED?	YES ☐ NO
ERIC REPRODUCTION RELEASE?	YES ☒ NO

001

100 TITLE
EVALUATION OF INDEXING BY GROUP CONSENSUS

101
102
103 FINAL REPORT

PERSONAL AUTHOR(S)

200 SCHULTZ, CLAIRE K. AND OTHERS

INSTITUTION (SOURCE) — SOURCE COD

300 INSTITUTE FOR THE ADVANCEMENT OF MEDICAL COMMUNICATION, PHILA., PA.

310 REPORT/SERIES NO.

OTHER SOURCE — SOURCE COD

320

330 OTHER REPORT NO.

OTHER SOURCE — SOURCE COD

340

350 OTHER REPORT NO.

400 PUB'L. DATE AUGUST -30, -1968 | CONTRACT/GRANT NUMBER OEC 1-7-070622-3890

PAGINATION, ETC.

500 33 PAGES
501

RETRIEVAL TERMS

600	INDEXING	METHODS
601	INDEXERS	RESEARCH
602	INDEXES	CRITERIA
603	EVALUATION	
604		
605		
606		

IDENTIFIERS

607 CRITERION GROUP METHOD

ABSTRACT

800 THIS METHOD TESTS THE EFFECTIVENESS OF TEST INDEXING SETS, USING A CRITERION
801 GROUP TO SET THE STANDARD FOR "IDEAL" INDEXING. THE CRITERION GROUP FOR A PARTI-
802 CULAR APPLICATION IS CHOSEN BY THE TEST ADMINISTRATOR, CONSISTENT WITH HIS OWN
803 CONCEPT OF WHO REPRESENTS THIS "IDEAL". MATCHING TEST SETS OF INDEXING TERMS WIT
804 THE CRITERION SET YIELDS AS MANY DEGREES OF MATCH AS THERE ARE MEMBERS OF THE
805 CRITERION GROUP (REFERRED TO AS CONCENSUS NUMBER). IMPORTANT VARIABLES FOR THE
806 METHOD ARE: SIZE OF DOCUMENT SAMPLE, SIZE OF CRITERION GROUP, INDEXERS' INSTRUC-
807 TIONS, METHOD OF EDITING RAW INDICIA TO MAKE THEM COMPARABLE, AND METHOD OF
808 WEIGHTING TERM SETS FOR SCORING.
809
810 RESULTS OF TESTING THE METHODOLOGIC VARIABLES FOR THEIR EFFECTS ON RELIABIL-
811 ITY, SENSITIVITY, FLEXIBILITY AND PRACTICALITY OF THE METHOD SHOW THAT "INDICATIV
812 TESTS CAN BE MADE AT THE 80% LEVEL OF CONFIDENCE WITH DOCUMENT SAMPLES AS SMALL A
813 10 AND CRITERION GROUPS AS SMALL AS 4; 95% CONFIDENCE REQUIRED COMPARABLE VALUES
814 AS LARGE AS 20 DOCUMENTS AND 9 CRITERION GROUP MEMBERS. THE 3 EDITING METHODS
815 TESTED: "NONE", MANUAL, AND COMPUTER, YIELDED DIFFERENT SCORES, BUT EACH PRESER-
816 VED DIFFERENCES BETWEEN TEST SETS, SO WAS NOT IMPORTANT TO SENSITIVITY OR RELIA-
817 BILITY, ONLY PRACTICALITY. SCORES CHANGED WITH DIFFERENCES IN INDEXER INSTRUCTIO
818 OR ADDITION OF A VOCABULARY GUIDE, SO THE METHOD IS SENSITIVE TO SUCH DIFFERENCE
819 BETWEEN TESTS BUT CAN BE CARRIED OUT SUCCESSFULLY WITH ESSENTIALLY NO INDEXER IN-
820 STRUCTION AND NO GUIDE. CONCENSUS NUMBER WAS ALMOST AS USEFUL AS CONCENSUS NUMBE
821 SQUARED FOR DETECTING DIFFERENCE IN TEST SETS; BUT THE LATTER WEIGHTING EMPHASIZE
822 "RECALL" VALUE TO SOME EXTENT.